Maria Kunilovskaya

# Translationese indicators for human translation quality estimation (based on English-to-Russian translation of mass-media texts)

A thesis submitted for the degree of Doctor of Philosophy

Research Institute of Language Processing
The Research Group in Computational Linguistics

November 2022

This work or any part thereof has not previously been presented in any form to the University or to any other body whether for the purposes of assessment, publication or for any other purpose (unless otherwise indicated). Save for any express acknowledgments, references and/or bibliographies cited in the work, I confirm that the intellectual content of the work is the result of my own efforts and of no other person.

The right of Maria Kunilovskaya to be identified as author of this work is asserted in accordance with ss.77 and 78 of the Copyright, Designs and Patents Act 1988. At this date copyright is owned by the author.

Signature: . . . . . .

Date: . . . . . .

# Abstract

Human translation quality estimation is a relatively new and challenging area of research, because human translation quality is notoriously more subtle and subjective than machine translation. At the same time, human translation is routinely assessed by education and certification institutions, as well as at translation competitions. Do the quality labels and scores generated from real-life quality judgments align well with objective properties of translations? This thesis puts this question to a test using machine learning methods.

Conceptually, this research is built around a hypothesis that the amount of translationese manifested in a translation can be predictive of its quality. This assumption is often made in translation studies but has never been put to a rigorous empirical test. Exploiting translationese features for quality prediction can help identify quality-related trends in translational behaviour and provide data-driven insights into professionalism to improve training. Translationese is a document level property, which fits well with the concept of quality in translation studies. Translationese features are also more interpretable and can explain linguistic differences between human translation quality categories.

We investigated an extended set of Universal Dependencies-based morphosyntactic features, inspired by previous research in translationese, as well as two lexical feature sets capturing (i) collocational properties of translations, and (ii) ratios of vocabulary items in various frequency bands along with entropy scores from n-gram models. To test our features against a challenging competition in translationese classifications and in quality estimation tasks, the experiments were run on *tf-idf* features, *QuEst++* features and contextualised embeddings from a range of pre-trained language models, including the state-of-the-art multilingual solution for machine translation quality estimation. Our major focus was on document-level prediction, however, where the labels and features allowed, the experiments were extended to the sentence level.

The corpus used in this research includes English-to-Russian parallel subcorpora of student and professional translations of mass-media texts, and a register-comparable corpus of non-translations in the target language. Quality labels for various subsets of student translations come from a number of real-life settings: translation competitions, graded

student translations, error annotations and direct assessment. We overview approaches to benchmarking quality in translation and provide a detailed description of our own annotation experiments.

Of the three proposed translationese feature sets, morphosyntactic features, returned the best results on all tasks. In many settings they were secondary only to contextualised embeddings. At the same time, performance on various representations was contingent on the type of quality captured by quality labels/score. Using the outcomes of machine learning experiments and feature analysis, we established that translationese properties of translations were not equality reflected by various labels and scores. For example, professionalism was much less related to translationese than expected. Labels from document-level holistic assessment demonstrated maximum support for our hypothesis: lower-ranking translations clearly exhibited more translationese. They bore more traces of passive translational behaviours associated with following source language patterns whenever possible, which led to the inflated frequencies of analytical passives, modal predicates, verbal forms, especially copula verbs and verbs in the finite form. As expected, they were more repetitive and had longer, more complex sentences. Higher-ranking translations are indicative of greater skill in recognising and counteracting translationese tendencies. For this quality type, translationese indicators might provide a valuable contribution to an effective quality estimation pipeline.

However, error-based scores, and especially scores from sentence-level direct assessment, proved to be much less motivated by translationese and fluency issues, in general. This is confirmed by relatively low regression results across all representations that had access only to the target language side of the dataset, by feature analysis and correlation between error-based scores and scores from direct assessment.

# Contents

**Abstract** iii

**List of Tables** viii

**List of Figures** xi

**Abbreviations** xiii

**1 Introduction** 1

1.1 Motivation and Background . . . 1

1.2 Research Aim and Objectives . . . 7

1.3 Research Design . . . 8

1.4 Main Contributions . . . 13

1.5 Publications . . . 14

1.6 Thesis Structure . . . 15

**2 Translationese Indicators** 17

2.1 Linguistic Specificity of Translations . . . 18

2.1.1 Key Concepts . . . 18

2.1.2 Features of Translation . . . 22

2.2 Translationese and Quality . . . 33

2.3 Proposed Features and Representations . . . 38

2.3.1 Feature Selection . . . 38

2.3.2 Universal Dependencies-based Features . . . 42

2.3.3 Collocations . . . 50

2.3.4 N-gram Language Modelling Features . . . 54

2.4 Summary . . . 55

**3 Human Translation Varieties** 57

3.1 Competence Levels . . . 57

3.1.1 Learner Translator Corpora . . . 58

3.1.2 Professionalism and Quality . . . 61

3.1.3 Expected Target Language Norm . . . 67

3.2 Data Description . . . 71
3.2.1 RusLTC Subsets . . . 71
3.2.2 Professional Translations . . . 77
3.2.3 Non-translations Corpora and other Resources . . . 78
3.2.4 Preprocessing . . . 80
3.3 Summary . . . 81
**4 Quality in Translation** **83**
4.1 Quality and its Aspects . . . 84
4.1.1 Theoretical Underpinnings of Quality . . . 85
4.1.2 Humans vs Machines . . . 87
4.2 Quality Estimation for MT and HT . . . 92
4.2.1 QE as NLP Task . . . 92
4.2.2 Representations in MTQE . . . 96
4.2.3 Human Translation Quality Estimation . . . 102
4.3 TQA and Designing Quality Benchmarks . . . 106
4.3.1 Quality Assessment and Benchmarking in HT and MT . . . 106
4.3.2 Reliability and Validity . . . 117
4.4 Quality Judgments for Student Translations . . . 121
4.4.1 Binary Quality Labels and their Verification . . . 122
4.4.2 Error-based Quality Scores . . . 125
4.4.3 Direct Assessment . . . 133
4.5 Summary . . . 138
**5 Translation Detection** **141**
5.1 Experimental Setup . . . 141
5.2 Translationese Classifications . . . 147
5.3 Feature Analysis . . . 154
5.3.1 Best of UD features . . . 154
5.3.2 Best of collgram features . . . 168
5.3.3 Best of n-gram features . . . 169
5.4 Discussion . . . 171

**6 Experiments on Varieties and Quality** **177**
6.1 Classification Tasks . . . . . 178
6.1.1 Translation Varieties . . . . . 178
6.1.2 Binary Quality . . . . . 183
6.2 Regression Tasks . . . . . 190
6.2.1 Experimental Setup . . . . . 190
6.2.2 Document Level: Error-based Scores . . . . . 192
6.2.3 Document Level: Direct Assessment . . . . . 198
6.2.4 Sentence Level: Error-based and DA Scores . . . . . 205
6.3 Discussion . . . . . 209
**7 Conclusion** **218**
**References** **234**
**Appendix A UD features by shorthand code name** **268**
**Appendix B Semantic groups of discourse markers** **280**
**Appendix C Core MQM taxonomy** **281**
**Appendix D *RusLTC* error taxonomy** **282**
**Appendix E Performance of neural classifiers** **285**
**Appendix F Comparative-contrastive UD features** **287**

# List of Tables


2.1 Examples of features proposed by Evert and Neumann (2017) . . . . . 26
2.2 Features associated with each translation trend in Volansky, Ordan and Wintner (2015) . . . . . 29
2.3 Types of linguistic information by language level captured with UD features 44

3.1 Parallel subcorpora of professional and student translations and non-translated reference subcorpus . . . . . 78
3.2 Language resources for collocational, LM and *QuEst++* features . . . . . 80

4.1 Translations by year of annotation . . . . . 127
4.2 Parallel subcorpora with error and DA annotation . . . . . 137

5.1 Professionals: translationese classification. Support: N(pro)=404, N(ref)=497149
5.2 Students: results of translationese classification. Support: N(stu)=334, N(ref)=497 . . . . . 149
5.3 Patterns in translational behaviour . . . . . 161
5.4 Properties of 28 UD features identified as discriminative between translations and non-translations by RFECV . . . . . 163
5.5 Additional corpus parameters for translationese classifications . . . . . 165
5.6 Properties of five and 14 UD features that were unique for professional and student experiments . . . . . 166

6.1 Classification result for translation professionalism labels (student vs professional) . . . . . 179
6.2 Professionalism: Results on pair-wise significance testing on SVM (left marker) and neural classifier (right marker) . . . . . 180
6.3 Basic parameters of parallel subcorpus with binary quality labels (after preprocessing and lemmatisation) . . . . . 183
6.4 Binary quality classification results (neural classifier) . . . . . 184

6.5 Quality labels: Results on pair-wise significance testing on SVM (left marker) and neural classifier (right marker) . . . . . 185
6.6 Features selected by RFECV with results of significance testing and comparison between quality categories . . . . . 188
6.7 Quantitative parameters of error-based scores across 553 documents . . . . 193
6.8 Regression results for error-based quality scores (553 documents) . . . . . 194
6.9 Correlation of averaged DA scores with quality aspects scores from error annotation (Spearman's r) . . . . . 199
6.10 Document-level: Parameters of DA scores distribution . . . . . 199
6.11 Document-level: Regression results for DA scores . . . . . 200
6.12 Regression results for 140 cross-annotated documents . . . . . 201
6.13 Regression results for random 140 documents with error-based scores . . . 202
6.14 Sentence-level results for unweighted error-based scores . . . . . 206
6.15 Sentence-level: Correlation of averaged DA scores with error-based scores . 208
6.16 Sentence-level: Regression results for DA scores . . . . . 208
6.17 Representation superior to UD on quality scores . . . . . 213

B.1 Semantic groups of listed discourse markers, including the number of items in each group by language and a few examples (ordered by frequency in the originally authored subcorpora) . . . . . 280

D.1 Distribution of error across categories . . . . . 282
D.2 Distribution of error attributes . . . . . 282
D.3 Weights assigned to error categories for accuracy and fluency scores . . . . 284

E.1 Professional translations: results of translationese classification (neural classifier) . . . . . 285
E.2 Student translations: results of translationese classification (neural classifier) 285
E.3 Student vs Professional translations (neural classifier) . . . . . 286
E.4 Bad vs Good: Binary quality classification (neural classifier) . . . . . 286

F.1 Properties of UD features sorted by degree of association between values in professional translations and non-translation . . . . . 287

F.2 Properties of UD features sorted by degree of association between values in student translations and non-translations . . . . . 289

# List of Figures


1.1 Size of student translations subsets with various types of quality judgments (number of sources, number of targets, number of aligned sentences) . . . . 10

3.1 Number of (sources, targets, aligned sentences) by type of annotation . . . 75

4.1 Annotation environment for binary labels verification . . . . . 123
4.2 Pairwise Krippendorff's alpha for re-assessment of documents with binary labels . . . . . 124
4.3 Percentages of major error categories to total errors . . . . . 126
4.4 Distribution of error tags across years of annotation . . . . . 128
4.5 Distribution of overall error-based quality scores for 553 documents . . . . 132
4.6 Document-level fluency scores for quality aspects . . . . . 132
4.7 Sentence-level weighted scores for quality aspects: accuracy and fluency . . 133
4.8 Sentence-level overall quality: unweighted and weighted . . . . . 133
4.9 Part of a bilingual document and the annotation task . . . . . 135
4.10 Distribution of scores from DA experiment for documents and sentences . . 138

5.1 Visualisations of PCA transform based on *UD* features . . . . . 151
5.2 Visualisations of PCA transform based on *mdeberta3* representation . . . . 152
5.3 More informative aspects of PCA analysis on other contextualised embeddings . . . . . 153
5.4 Weights loading the decision towards each category for 32 RFECV-selected items . . . . . 155
5.5 Weights loading the decision towards each category for 42 RFECV-selected items . . . . . 156

6.1 Variability of PCA D1 values in quality-annotated documents (along with their sources and non-translations) . . . . . 185
6.2 Variability of PCA D1 values in quality-annotated documents (no source texts) . . . . . 186

6.3 Weights loading the decision towards each category for 26 RFECV-selected items . . . . . 189
6.4 Distributions of accuracy scores for error-annotated and cross-annotated datasets . . . . . 202

D.1 Screenshot of a *brat*-based annotation environment . . . . . 283

# Abbreviations

**ANOVA** analysis of vaiance

**API** application programming interface

**ATA** American Translation Association

**AUC** area under (ROC) curve

**BERT** Bidirectional Encoder Representations from Transformers

**BiLSTM** bi-directional Long Short-Term Memory

**BLEU** Bilingual Evaluation Understudy

**BNC** British National Corpus

**CNN** convolutional neural networks

**DA** Direct Assessment

**DM** discourse markers

**DQF** TAUS Dynamic Quality Framework

**EWT** English Web Treebank

**FTD** Functional Text Dimensions

**HT** human translation

**HTER** Human Translation Error Rate

**HTQE** human translation quality estimation

**IRA** Inter-rater agreement

**IRR** Inter-rater reliability

**L1** mother tongue

**LDA** Linear Discriminant Analysis

**LM** language model

**LTC** learner translation corpora

**MAE** Mean Absolute Error

**METEOR** Metric for Evaluation of Translation with Explicit ORdering

**MFW** most frequent words

**MI** mutual information

**ML** machine learning

**MQM** Multidimensional Quality Metrics

**MT** machine translation

**MTQE** machine translation quality estimation

**MWE** multi-word expressions

**NLP** Natural Language Processing

**NMT** Neural Machine Translation

**NPMI** normalised pointwise mutual information

**OOV** out-of-vocabulary

**PCA** Principal Component Analysis

**PE** post-editing

**PMI** pointwise mutual information

**PoS** Parts of speech

**QE** quality estimation

**ReLU** Rectified Linear Units

**RFE** recursive feature elimination

**RFECV** Cross-validated RFE

**RMSE** Root Mean Square Error

**RNC** Russian National Corpus

**RNN** recurrent neural networks

**ROC** receiver operating characteristic

**RusLTC** Russian Learner Translator Corpus

**SL** source language

**SMT** Statistical Machine Translation

**SOTA** state-of-the-art

**ST** source text

**STD** standard deviation

**STS** Semantic Textual Similarity

**SVM** Support Vector Machines

**SVR** Support Vector Regressor

**tf-idf** term frequency–inverse document frequency

**TL** target language

**TMX** Translation Memory eXchange

**TQA** Translation Quality Assessment

**TQE** Translation Quality Estimation

**TS** Translation Studies

**TT** target text

**TTR** type-to-token ratio

**UD** Universal Dependencies

**WMT** Conference on Machine Translation

**XLM-R** XLM-RoBERTa

**XML** Extensible Markup Language

# Chapter 1 Introduction

## 1.1 Motivation and Background

My experience in translator education of many years led me to believe that translations, and student translations in particular, share specific features that make them recognisable as such. Year after year, I annotated and commented the same suboptimal translation solutions. Their repetitiveness suggested that student translations can have automatically recognisable properties, which can be used to predict their quality.

This thesis presents a comprehensive study of the relations between translationese and translation quality as captures by various assessment methods. It explores claims about translationese in English-to-Russian translation made in practical translation textbooks and arising from personal experience, and extends the existing scope of knowledge about the properties of translated language, making them even more visible to a professional eye. It often requires to know where to look to actually see. Raising awareness of the peculiarities of translated language can be useful in translator training, helping students to avoid typical pitfalls and recognise the cognitive constraints of the translation process. Importantly, this thesis investigates the extent to which translationese properties are taken into account in judgments about translation quality.

**Area of study and rationale** We draw from several research fields at various stages of this project trying to bridge the gap between theoretical translation studies and supervised machine learning (ML) methodologies. They can be roughly outlined as follows:

- translation quality related theories, including assessment approaches;
- empirical and computational translation studies focused on translationese research and studies on professionalism in translation;
- corpus linguistics focused on student translation and (register-)comparable corpora;
- machine translation quality research including human assessment methods and quality benchmarks;
- machine learning approaches to solve translationese-related and quality estimation

tasks.

The specificity of translations in comparison with their sources and non-translated texts in the same target language register was acknowledged in literary criticism long ago. For example, Berman formulated 12 deforming tendencies in literary translations, including rationalisation (making text more coherent), ennoblement (more elegant style), destruction of underlying networks of signification, etc (Berman, 1985). However, it was not before Baker's seminal work (Baker, 1993) that translations became an object of study as a special type of text, and the term *translationese* took root. After almost three decades of research, translationese can be broadly understood as statistical deviations of translations from the expected target language norm manifested in comparable originally-authored texts in the target language (TL). The body of research exploring peculiar regularities of translated language can be referred to as *translationese studies*.

From early works on, translationese was viewed as a strictly descriptive, non-evaluative term reflecting the nature of translations (Gellerstam, 1986; Baker, 1993). It was established that all translations are inevitably different from comparable originally-authored texts. It does not mean that translated texts are written to a lower professional standard than comparable non-translations (De Sutter et al., 2017). Instead, they form their own language subsystem and can be approached as a language variety.

At the same time, the amount and type of translationese observed in various categories of translations can be indicative of translation quality. For translation scholars and educators, professionalism in translation is often associated with the ability to identify translation problems and find the most acceptable solution for them (Pym, 2003), i.e., most importantly, to be able to recognise and counteract the natural pull of the source language which triggers literal renditions.

The association between the amount of translationese and translation quality was either assumed or demonstrated in a number of publications, such as Scarpa (2006); Rabadán, Labrador and Ramón (2009); De Sutter et al. (2017) as well as in Loock, Mariaule and Oster (2013); Loock (2016) (as quoted by De Sutter et al., 2017). For example, Scarpa (2006) found that simplification correlated with lower-scoring translations and explicitation correlated with higher quality scores. De Sutter et al. (2017) interpreted the distance between individual student translations and comparable professional texts in scatter plots

as indicative of lower quality. They also found statistical differences between student and professional translations in a univariate analysis on a number of language-independent easily-extractable features in English-to-French translation of fiction and in French-to-Dutch news. Our own early univariate exploration of student translations indicated that they were more distant from the reference corpus than professional translations on a number of variables (Kunilovskaya, Morgoun and Pariy, 2018).

Despite *translationese studies* is an established research direction in the empirical translation studies, the applicability of translationese features to human translation quality estimation (HTQE) remains unclear. It is especially surprising, given that translationese-based approach was used in a much more dynamic and technologically advanced area of machine translation quality estimation (MTQE) (Aharoni, 2015) and in the machine translation detection task (Carter and Inkpen, 2012). In particular, Aharoni (2015) showed that the higher the translationese classification accuracy (i.e. the better machine translations were distinguished from non-translations), the lower the expected translation quality. Aranberri (2020) also found translationese predictive of machine translation (MT) quality. Interestingly, Rubino, Lapshinova-Koltunski and Van Genabith (2016) used surprisal and distortion measures derived from delexicalised language models which they refer to as quality estimation indicators to discriminate between novice and professional translations. Nonetheless, we are not aware of any empirical studies that explore the suggested feasibility of a translationese-based human translation quality model, except our own previous work (Kunilovskaya and Lapshinova-Koltunski, 2019, 2020).

**Hypothesis** Thus, this research is designed to test a hypothesis that translationese indicators are predictive of human translation quality. This hypothesis will be confirmed if we find that lower-ranking translations exhibit stronger translationese tendencies that their higher-ranking counterparts.

Generally, this research contributes to a fairly new research area of automatic prediction of human translation quality. The term HTQE seems to have first appeared in a PhD thesis by Yuan (2018) and a subsequent publication by the same author (Yuan and Sharoff, 2020). In their work, the concept of quality estimation for MT is adopted to predict English-to-Chinese human translation quality. The alternative automatic approach to MT quality (quality evaluation) measures deviations from a MT candidate and

human reference translations. It is hardly applicable to human translation even theoretically due to a wide range of legitimate variation in linguistic expression that translators can resort to. In human translation, it seems unfair to penalise renditions because they deviate from a pre-selected good translation. Zhou and Bollegala (2019) formulated the task of predicting human translation quality as an unsupervised graph matching problem. They employ cross-lingual word embeddings to represent sources and targets, and measured the distance between the respective vectors, reporting Spearman 0.529 for their Japanese-to-English data.

Other attempts to develop automatic approach to measure human translation quality were not formulated as a ML task, but present corpus-based studies and observations. De Sutter et al. (2017) report results of a multifactorial corpus analysis backed with Principal Component Analysis (PCA)-based visualisation to give a translation teacher a method to 'objectify formative quality assessment'. They perform monolingual comparison, presenting student translations against the backdrop of comparable professional translations and non-translations as two possible TL norms, and reducing the concept of human translation quality to its acceptability aspect.

**Anticipated issues and challenges** We envisage a few theoretical and practical challenges that should be kept in mind. This paragraph discusses caveats that limit the scope of this research and possible contingency measures to curb their effects on the outcomes. One limitation of a translationese-based approach to translation quality estimation is that it reflects only one of translation quality aspects, namely fluency (also referred to as acceptability/readability). It does not have access of the semantic relation between source and target texts, known as the accuracy/adequacy aspect of translation. Nonetheless, as it is shown in Sections 2.2 and 6.1.2, there are good reasons to believe that

- translationese-related fluency might have a large impact on human perception ('It does not sound right reaction'),
- fluency and accuracy might be difficult distinguish in an annotation task,
- fluency and accuracy might correlate to a great extent – it is difficult to imagine a human translation produced in good faith that is very fluent but inaccurate; inaccuracies usually render translated text less readable and coherent than expected and are rarely seamless,

- differences between student and professional translations are mostly related to the fluency aspect of translation quality.

Frequency-based translationese features is not immune to sparsity. Translationese-based approach is less applicable to the sentence-level quality estimation.

Assessing translations on the basis of translationese indicators, requires carefully selected external resources that would represent the expected standard, i.e. either comparable professional translations, or reference texts, or both. It is not uncommon to collect and maintain such resources in both educational (Bowker, 2001) and industrial settings (translation memories and TL reference corpora) (Massey and Wieder, 2019), but clearly such resources can be less available in other circumstances. Constructing a comparable corpus of non-translations to represent TL textual fit (Chesterman, 2004) is one of the challenges in this project. Register and translational conventions in each register are known to be one on the major factors shaping linguistic properties of translations (Lapshinova-Koltunski, 2017). A preliminary step in this research was to design an approach to select a subset of non-translation from a large TL corpus that would be functionally-comparable to the English source texts in the parallel subcorpora (Kunilovskaya, Taslimipoor and Ilyushchenya, 2019).

By design, this research is focused on shining-through form of translationese. Some researchers express reservations whether SL/TL-induced translationese can explain all specificity of translations. They name other factors that might be at work, including translators' awareness of their sociocultural roles and positions, i.e. translation norms (Laviosa, 2008; Delaere, De Sutter and Plevoets, 2012), extralinguistic conditions of the translation situation such as time constrains and perceived responsibility, rather than lack of professionalism and inadequate language use. It is necessary to control for some of these other factors in the experiment design by selecting register-homogeneous translations produced under similar communicative and translation conditions by many subjects of known professional status.

**Practical relevance** This thesis primarily contributes to the understanding of translationese and its relation to quality judgments. Our results can be used as a module in a HTQE tool to estimate the translations quality with regard to the amount of translationese manifested it a translated document. There is a number of the applications for

HTQE, primarily in education, certification and quality control.

Probably the largest sphere where there is a wide need to score human translations is translator education and language learning scenarios. Translation programmes of various standing (university degrees in translation, vocational training courses or advanced professional development programmes) as well as practical translation courses included into the foreign language education programmes routinely grade students' written submissions. Some form of summative assessment is indispensable in the test and exam settings. However, despite institutional grading manuals can be very detailed, translation quality assessment often proceeds "according to the lordly, but completely unexplained, whimsy of 'It doesn't sound right.'", a much quoted and accurate description offered by Fawcett (1981, p. 142). An automated approach can make this process more objective and systematic, devoid of personal preferences and expectations, more uniform in terms of rigour, and immune to the known negative effects common for all human evaluation, including lapses of attention and the order of translations to be graded (Minacori Vibert, 2010).

A reliable automation of the scoring process can help to better monitor translation students' progress and provide qualitative feedback on their work more often. It is true that getting regular formative human feedback on their course work between the tests would be more useful, but, given the class sizes and the enormous effort required to assess translations manually, it is unavailable in many translation programmes. Automatic feedback can also contribute to self-evaluation and independent learning.

Apart from qualification exams in the training environments, many countries have certification requirements that facilitate the access to professional jobs. Certification usually includes a translation task, which is scored by several trained evaluators. In the context of the American Translation Association (ATA), there is a points-based threshold (more than 46) after which the error annotation can be abandoned because the translation is obviously below the expected quality standard[1]. An automatic scoring algorithm can be used as a preliminary filter to identify the submissions that can be desk-rejected for being significantly below the standard. A similar functionality can be useful for filtering out very poor submissions to translation contests to help a human jury focus on, for example, the top 30% of the participants to identify the winners.

In the industrial contexts, an automatic scoring tool for human translations can be

[1]see the instructions at www.atanet.org/certification/Framework_2017.pdf

applied on both the language service provider side for internal quality control purposes or on the customer side for ensuring the quality of the translation obtained from a translator or a translation agency. It can be especially useful if the translation user does not possess a bilingual competence and does not want to involve a third party for a linguistic expertise. A similar need for a quick and cheap evaluation of translation quality arises when several human translations of the same sample are available and the customer is seeking to detect the best of them.

## 1.2 Research Aim and Objectives

This thesis sets out to explore the performance of translationese indicators inspired by contrastive linguistics and translation studies in the task of predicting a range of available human judgments.

To achieve this overarching purpose, we define a number of research objectives:

1. Develop a set of translationese indicators motivated by previous findings in translationese studies and contrastive analysis for English-to-Russian language pair (mass-media register).
2. Test the proposed feature sets in a translation detection task (translationese classification) on two competence-based varieties (student and professional translations). This objective includes:
   (a) exploring ways to build a motivated comparable corpus of non-translations, the second class in the translationese-diagnostic binary classification;
   (b) comparing the performance of the proposed feature sets to each other and to the alternative representations, including multilingual/cross-lingual embeddings;
   (c) identifying and describing the strong translationese indicators.
3. Collect various types of human quality judgments. This involves a number of preliminary research tasks and practical steps:
   (a) explore the approaches to produce quality labels in human and machine translation applicable to our data;
   (b) build a comparable parallel corpus of professional translations and generate a reasonably diverse subset of student translation from an existing learner

translation corpora (LTC) to represent competence-based categories as a proxy of quality;

(c) produce quality labels from Russian Learner Translator Corpus (RusLTC) 'grade' attribute and continuous document- and sentence-level scores from existing error annotation.

(d) set up a direct assessment annotation experiment and produce respective document- and sentence-level scores.

4. Explore alternative approaches to text representation to predict document- and sentence-level quality scores. More specifically:

   (a) use *tf-idf* as a simple baseline,

   (b) generate alternative explicit features designed for MTQE within QuEst++ framework,

   (c) apply cross-lingual sentence encoders, fine-tuned on Semantic Textual Similarity (STS) task and on DA scores in a MTQE task,

   (d) generate representation from state-of-the-art general-purpose word embedding models: a dedicated Russian language model and a multilingual *mdeberta3-base*.

5. Compare the outcomes of classification and regression experiments, supported by feature analysis where possible, across various representations and types of quality judgments.

## 1.3 Research Design

This research focuses two components of a quality estimation task: types of quality labels/scores and text representations. We do not lay much emphasis on the learning algorithm and use a simple Support Vector Machines (SVM) and a one-layer neural model for sanity.

**Quality labels** One of the important questions that we explore in this thesis is generating benchmarks, i.e. sets of gold-standard labels/scores reflecting human judgments

about translation quality that can be used as learning targets in a ML setup. Weak artificial intelligence, as a research and engineering domain, seeks to teach the machine to solve tasks that can be performed by humans. For now, judging human translation quality is perceived as one of the tasks that is challenging for the machines. It is not entirely because it requires unstructured professional knowledge to yield to automation. It is also because translation quality is difficult to capture in a reliable way due to the notorious subjectivity of the underlying concept. This thesis compares the performance of various document representations, including translationese indicators, on labels/scores from several methods to annotate human translations for quality. These experiments can reveal dissimilarities between assessment methods with regard to the type of quality they capture. Our experiments are run on the four types of quality categories or judgments:

1. Binary quality: coarse-grained binary quality labels, derived from holistic assessment of exam translations in several Russian universities and of submissions to a number of translation competitions;
2. Professionalism: ontological status of texts based on the meta information about translators;
3. Scores from error annotation: aggregated measures of quality based on error statistics;
4. Scores from direct assessment: a mean score across two annotators.

Annotated student translations come from an existing multi-parallel LTC which collected quality assessment carried out in real-life contexts (student translation competitions, graded exam translations, error annotation as part of the formative feedback). The quality judgments were produced outside of a well-controlled – and artificial – experimental setup. This explains the lack of all four labels for all student translations in our research corpus. Instead, we have multiple intersecting subsets shown in Figure 1.1. We tried to acquire as much corpus data for each type of assessment as possible without compromising the integrity of the categories.

A detailed description of how each subset was selected can be found in Section 3.2.1; the annotation setups and associated reliability studies appear in Section 4.4.

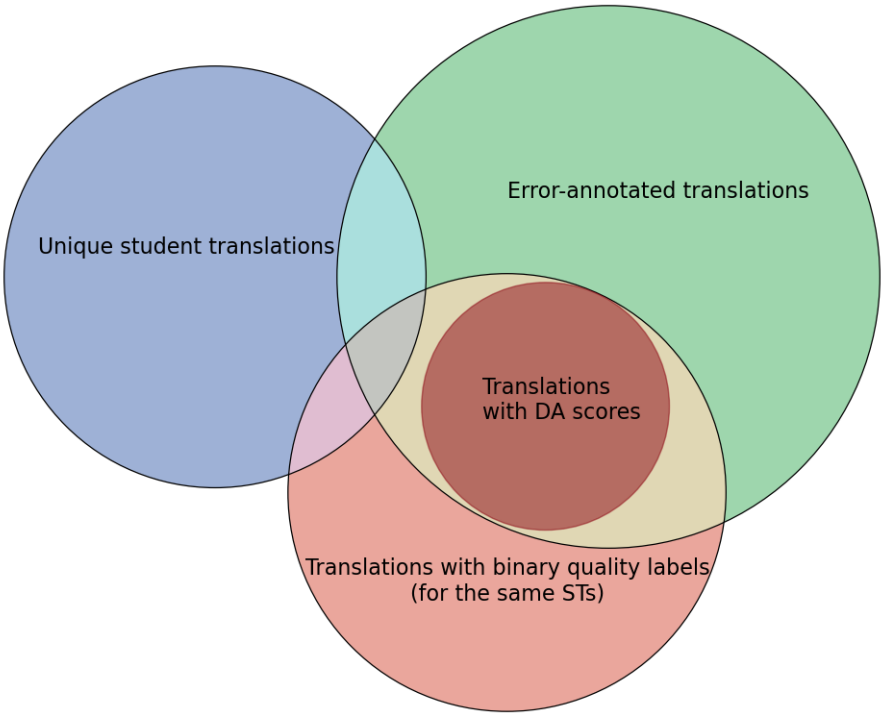



Figure 1.1: Size of student translations subsets with various types of quality judgments (number of sources, number of targets, number of aligned sentences)

**Features and representations** In this thesis the performance of several sets of well-established and novel language-pair specific translationese indicators is thrown into perspective of bilingual features from *QuEst++*[2], a translation quality estimation framework for MT (Specia, Paetzold and Scarton, 2015), and a number of contextualised embedding models.

Traditionally, the explorations of translationese are based on manually-selected explicit features that capture a particular phenomenon represented by its normalised frequency. Probably the most popular feature set of diverse and elaborate hand-engineered features was proposed by Volansky (2012), and later extended by Sominsky and Wintner (2019). In the last decade the research paradigms shifted from univariate statistical analyses to supervised and unsupervised ML methods, which take into account feature combinations. A standard practice is a bottom-up approach which includes as many features as possible

[2] github.com/ghpaetzold/questplusplus

and empirically establishes their effectiveness through feature selection (Evert and Neumann, 2017; Yuan, 2018). An alternative approach consists in producing sparse or dense document vectors from sequences of items in that document. Vectorisers were run on character, word or Parts of speech (PoS) n-grams (, respectively Popescu, 2011; Bernardini and Baroni, 2005; Carter and Inkpen, 2012; Eetemadi and Toutanova, 2015; Rabinovich, Ordan and Wintner, 2017; Lapshinova-Koltunski, 2019). More recent studies make use of language modelling approaches (i) to measure entropy/perplexity of a model trained on non-translations and run on translations (Karakanta and Teich, 2019; Nikolaev et al., 2020); or (ii) to generate static embeddings for PoS or semantic tags sequences in translated and non-translated corpora (Chowdhury et al., 2020; Chowdhury, Espãna-Bonet and van Genabith, 2021). This thesis compares the performance of explicit linguistically-motivated features against representations from a number of embedding models, which are supposed to capture various properties of translations. The rationale behind using embeddings is that these representations might capture complex non-local syntactic and semantic information that is difficult to encode manually. Neural representations are expected to perform better than hand-engineered features, but relative ease of generating embeddings and expected superior performance come at the expense of explainability. At the moment there is no published work on translationese (or HTQE) using recent contextualised embedding models, except Pylypenko et al. (2021). In this thesis we compare the performance of hand-crafted translationese indicators and document-averaged embeddings to solve the tasks of translation detection and HTQE.

Although this research is based on human translations, we make a point of exploring theoretical and practical approaches used in MTQE. We are aware of the ontological differences that make these two translation varieties very dissimilar, but we hope that there can be a productive interconnection between the two closely related research areas: machine translation quality estimation and quality assessment in translation studies. We see it as part of our task to reveal the opportunities for mutually beneficial interface between machine translation and translation studies. Given our primary focus on student translations, we are mostly interested in borrowing from MT approaches to solve our task. At the same time, the increased quality of MT, especially with regard to fluency, makes the task of MTQE more demanding in terms of ability to reveal fine-grained quality distinctions. Besides, MT paradigm can benefit from shifting to document level, which

is the default setting in translation studies, and is called for in MT, especially when comparing its performance to human translators (Läubli, Sennrich and Volk, 2018; Voita, Sennrich and Titov, 2019).

The particulars of a research design depend on the desired trade-off between interpretability, scalability and performance, among other factors. Our preferences and interests are more on the translation studies side: we use ML methods to understand the nature of translationese better and to produce interpretable automatic quality predictions, if possible. Previous research returned mixed results on language-independent generic features (De Sutter et al., 2017; Kunilovskaya, Lapshinova-koltunski and Mitkov, 2021*b*). Some researchers found that language-specific indicators return better results in translationese studies (Hu and Kübler, 2021). We are prepared to sacrifice some scalability of the approach for higher explaining power, validity and diagnostic potential. To this end, we engineer features, based on theoretical or empirical evidence from translation studies or contrastive linguistics for a given translation direction. The use of embeddings is considered here mostly as a means to demonstrate sanity of the quality annotation and as a way to throw our feature-based models into perspective of the achievable for each type of labels/scores.

**Document- vs sentence-level** This research is focused on the document-level scenario as it is more in line with the nature of human translation. In most real-life scenarios, human translation evaluation is required at document level. However, it can be helpful if an algorithm could highlight segments that might require reworking or additional editing due to being inaccurate with regard to the source or too translational at the level of expression. Besides, sentence- or segment-level predictions are more compatible with neural models. *Longformer* approaches (Beltagy, Peters and Cohan, 2020) are still in the making and are not available for Russian at the time of writing, while representing texts as an aggregation of sentence vectors might be a very crude simplification.

Our experiments are framed as text classification/regression problems solved with the default linear-kernel SVM (and a simple one-layer neural model for comparison and sanity check). We use several feature selection strategies to reveal and describe the best-performing features.

## 1.4 Main Contributions

We see the main contributions of this project as follows:

- We developed a theoretically-motivated feature set for translationese diagnostics in English-to-Russian translation. Its applicability to other language pairs (with adaptation) and registers was tested in other projects, but these results are outside the scope of this thesis.
- We revealed strong translationese indicators, which were mostly associated with the *shining-through* trend in translational behaviour, and demonstrated that lower-ranking translations exhibited more translationese than higher-ranking translations.
- This thesis showed that morphosyntactic translationese indicators are competitive against QuEst++ features and most less interpretable distributed representations, at least on document-level labels from holistic assessment.
- We established that distinctions between professional and student translations were less related to translationese than expected. Professional translators might be better at recognising and counteracting some influences of the ST, but they generate other types of translationese which makes them easy to distinguish from non-translations. If anything, professional translations are more homogenous than student translations, probably due to a more standard repertoire of translation solutions.
- We revealed dissimilarities between three quality assessment methods in terms of sensitivity to translationese and in terms of capturing document-level properties. Document-level holistic assessment was most translationese aware, error-based scores demonstrated some alignment with translationese properties of texts and performed better at document-level. Scores from direct assessment returned better results at sentence level and were largely agnostic of translationese.
- We released datasets for document- and sentence-level HTQE experiments in English-to-Russian language pair with three types of quality judgments[3]. First, there is a set of 360 documents with good/bad labels. Second, there is a Conference on Machine Translation (WMT)-format dataset of 553 documents/12,369 sentences with error-based scores, including 140 documents/3,224 sentences annotated in DA setup.

[3]The datasets are available from the RusLTC website: www.rus-ltc.org/static/html/about.html

## 1.5 Publications

Preliminary results related to the ideas developed in this thesis appeared in the following publications:

1. Kunilovskaya, M. and Sharoff, S. (2019) 'Towards functionally similar corpus resources for translation', *Proceedings of Recent Advances in Natural Language Processing*, Varna, 2-4 September, pp. 583–592, doi: 10.26615/978-954-452-056-4_069.
2. Kunilovskaya, M., Taslimipoor, Sh. and Ilyushchenya, T. (2019) 'Functional text representations for building cross-linguistic comparable corpora in English and Russian', *Proceedings of 12th Workshop on Building and Using Comparable Corpora (BUCC 2019)*, Varna, 5 September, pp. 37–45. Available at: comparable.limsi.fr/bucc2019/Kunilovskaya_BUCC2019_paper5.pdf (Accessed: 16 June 2022).
3. Kunilovskaya, M. and Lapshinova-Koltunski, E. (2019) 'Translationese features as indicators of quality in English-Russian human translation', *Proceedings of the 2nd Workshop on Human-Informed Translation and Interpreting Technology (HiT-IT 2019)* (pp. 47–56), Varna, 5-6 September, doi: 10.26615/issn.2683-0078.2019_006.
4. Kunilovskaya, M. and Lapshinova-Koltunski, E. (2020) 'Lexicogrammatic translationese across two targets and competence levels', *Proceedings of the 12th Conference on Language Resources and Evaluation*, Marseille (online), 13–15 June, pp. 4102–4112. Available at: aclanthology.org/2020.lrec-1.505 (Accessed: 16 June 2022).
5. Kunilovskaya, M. and G. Corpas Pastor (2021) 'Translationese and register variation in English-to-Russian professional translation', in Lim, L. and Li, D. and Wang, V. (eds.) *New Perspectives on Corpus Translation Studies*. Singapore: Springer, pp. 133–180, doi: 10.1007/978-981-16-4918-9_6.
6. Kunilovskaya, M., Ilyushchenya, T., Morgoun, N. and Mitkov, R. (2021) 'Source language difficulties in learner translation: evidence from an error-annotated corpus', *Target*, doi: 10.1075/target.20189.kun.
7. Kunilovskaya, M., Lapshinova-Koltunski, E. and Mitkov, R. (2021) 'Translationese in Russian literary texts', *Proceedings of the 5th Joint SIGHUM Workshop on Com-*

*putational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature at EMNLP*, Punta Cana (online), 11 November, pp. 101–112, doi: 10.18653/v1/2021.latechclfl-1.12.

Some of these papers report results for similar experimental setups. However, the results reported in this thesis are based on a fresh run of these experiments after we had addressed the issues identified in preliminary studies.

## 1.6 Thesis Structure

In brief, this thesis is structured around the research steps. Chapters 2-4 discuss previous research relevant for each project's task and describe proposed solutions. Chapters 5-6 describe the experimental setup, report and analyse the outcomes of the experiments.

In particular, Chapter 2 introduces translationese studies, a research direction within empirical translation studies, which describes and explains linguistic properties of translations. We provide the main theoretical underpinnings of our research hypothesis, showing the link between translationese and translation quality. The last part of the chapter gives an overview of the proposed feature sets along with the extraction procedures. We also explain the rationale behind feature selection, using existing empirical evidence and approaches to translation detection.

Chapter 3 discusses previous research on differences between competence-based (professional) varieties of translated language (student and professional translations), which are often interpreted as a proxy for quality. This includes a brief overview of learner translator corpora as a common source of human quality-annotated translations. A separate paragraph deals with theoretical and practical issues of building cross-linguistically comparable corpora for translationese research. The second part of the chapter gives more details on the sources and nature of the textual data in this research. It presents the rationale for subsetting student translations from an existing LTC. The description of textual data, including large external resources required for feature extraction, is complete with quantitative parameters and preprocessing details.

Chapter 4 compares MT and human translation (HT) in the context of quality: we give a concise overview of the main theoretical aspects of translation quality and summarise

approaches to quality estimation, including previous research in HTQE. The chapter focuses text representations for QE task, highlighting the more prominent feature-based and embedding-based solutions. The second part of the chapter overviews established practices and recommendations to produce manual 'gold standards' for ML experiments modelling translation quality. Importantly, Chapter 4 contains a description and results of cross-annotation, error annotation and direct assessment experiments, which yielded quality labels/scores used in this thesis.

Chapter 5 presents the results of translationese classifications. We compare the performance of a number of representations on professional and student translations. A considerable part of the chapter is given to feature analysis using automatic feature selection and univariate correlation and statistical significance testing. It allow us to identify strong translationese indicators in English-to-Russian translation of mass-media texts.

Chapter 6 is our main experimental chapter. It opens with classification results on professionalism labels and binary quality labels. Section 6.2 reports regression outcomes for error-based and DA scores. The experiments are run at document and sentence levels. The results at document level are illuminated with feature analysis.

Chapter 7 summarises our findings and draws conclusions by chapter. We discuss limitations and possible directions for future work.

# Chapter 2 Translationese Indicators

This thesis explores the relation between theoretically-motivated and interpretable translationese features and available quality judgments on student translations. We also investigate translationese-related distinctions between professional varieties of translations. This research goal makes feature engineering and selection a cornerstone of the project and necessitates an overview of translationese-related studies with a focus on recent developments in feature-based approaches and related research methods.

The term *translationese* was introduced by Gellerstam (1986). A few year later, Baker (1993) convincingly called for translation scholars to shift attention from source text (ST)-target text (TT) relation, which traditionally cast translations as unworthy 'second-hand' texts, to objective properties of translations distinguishing them from non-translations in the TL, the phenomenon of translationese has become the one of the main objects of study in empirical (or corpus-based) translation studies. Today, the study of translationese is a well-established research field, which relies on computational methods. It gave rise to a number of interrelated Natural Language Processing (NLP) tasks such as translation detection, translation direction and SL identification. Machine learning methods are used to investigate variation in translation along a number of dimensions such as register, professional competence, source language, translation method (automatic vs. human), etc.

This chapter opens with the key theoretical concepts in the field and focuses on the more recent developments regarding feature engineering and text representation. In Section 2.2 we present theoretical underpinnings and existing evidence that substantiates our main hypothesis that the amount of translationese is correlated with translation quality. Section 2.3 has a detailed description of translationese feature sets proposed in this research.

## 2.1 Linguistic Specificity of Translations

### 2.1.1 Key Concepts

This thesis builds on the observation that translated texts have regularities in their linguistic form making them distinct from texts generated outside the communicative situation of translation. These unique characteristics of translations are attributed to the specificity of the underlying communicative process. They are the 'fingerprints' of the translation process on the product (Gellerstam, 1986).

The observed peculiarities of translated language gave rise to a number of explanatory theories that focused various aspects of this phenomenon and suggested their own terms to refer to it. The major schools of thought can be distinguished based on whether the specificity of translations is attributed to (i) the factors, independent of the language pair involved, (ii) to the confrontation of (specific) source and target languages or (iii) mostly to the SL interference with some normalisation pull from the TL.

The first strand of research can be best represented by the 'translation universals' interpretation, put forward by Baker (1993). She argues that translations demonstrate properties that cannot be linked to respective SLs, but reflect the constraints of the translation process per se. According to Baker (1993, p.243), universal features of translation are "features which typically occur in translated text rather than original utterances and which are not the result of interference from specific linguistic systems". She discusses four hypotheses that might explain the specific pattering of translations (Baker, 1996, pp.176–177, 183–184):

- *simplification* (a subconscious tendency to simplify the language or message or both of the source),
- *explicitation* (a tendency to spell things out rather than leave them implicit),
- *normalisation/conservatism* (a tendency to conform to patterns which are typical of the target language) and
- *levelling-out* (a tendency to gravitate towards the center of a continuum).

Blum-Kulka, whose work originally introduced the explicitation hypothesis, Blum-Kulka (1986), made a conclusion that cohesive patterns observed in translations were neither SL nor TL oriented.

The second explanation is offered by some scholars who characterise the linguistic makeup of the translations as 'third language' (Duff, 1981), 'third code' (Frawley, 1984) or 'hybrid text' (Schäffner and Adab, 2001*a*,*b*). They tend to explain the distinctive features of translations by the confrontation of specific linguistic systems, which result in mixed 'strange' patterns of linguistic expression making translations their own TL subsystem.

The third approach attributes deviations of the translated language from the expected TL norm primarily to the SL influence. Gellerstam (1986) and Santos (1995), who use the term 'translationese' to refer to linguistic properties of translations, tend to consider source language influence the major cause of linguistic specificity of translations. For example, Santos (1995) uses the term 'translationese' "to denote the influence of properties of the source language in a translated text" (p.61). Lexical generalisations made by Gellerstam (1986) reflect the various degrees of impact of the SL, although he acknowledges that not all translationese is explainable by interference: cultural differences and cross-linguistic divergences between registers are other important sources of translationese. A more radical stance is taken by Rabadán, Labrador and Ramón (2009), who defines translationese as "features of translated language that can be attributed to the influence of the source language" (p.303).

Toury (1995) maintains a descriptive approach to the study of translated texts. He explains the specificity of translations through 'laws' (most importantly *the law of interference* and *the law of growing standardisation* also known as normalisation, i.e. a tendency to adopt to linguistic expression in translation to the prototypical features of the target language (Denturck, 2009)) and 'norms', i.e. socio-cultural constraints regulating the behaviour of professional translators. The latter can be exemplified by 'language prestige', an idea that an unequal status of languages and cultures could affect the tolerance of interference. This effect was demonstrated in a study of Finnish translations from English and Russian. Mauranen concluded that there was a greater tolerance in the Finnish culture for English than for Russian interference (Mauranen, 2004, p.72). Similarly, more SL-related translationese was observed in German translations that in English translations in a bidirectional study on English and German by Evert and Neumann (2017).

More recent translationese studies, including (Koppel and Ordan, 2011; Volansky,

Ordan and Wintner, 2015; Rabinovich, Ordan and Wintner, 2017; Chowdhury et al., 2020; Chowdhury, Espãna-Bonet and van Genabith, 2021; Kunilovskaya and Lapshinova-Koltunski, 2020) and (Evert and Neumann, 2017) above, demonstrated that a lot of translationese could actually be explained by SL *interference*, or *'shining through'*, a term introduced by Teich (2003). It is defined as "diverging frequencies of options existing in both languages are adapted in translated texts to those of the source language". Note that some authors use the term 'interference' interchangeably with 'transfer'. However, if the distinction is made, 'interference' stands to mean 'negative transfer', "a gross deviation from the target language norm, or what could be called a translation error in one sense", according to Mauranen (2004, p.71–72). Positive transfer is the reliance on the intersection of patterns found in both languages. In this work we do not focus on this distinction, although our features attempt to capture differences in the use of intersecting patterns.

Chesterman (2004) emphasised the influence of each member in the language pair by arranging the tendencies discussed in translationese studies into SL- and TL-universals depending on the type of comparison required to establish them. Some researchers describe linguistic properties of translations as 'split' between normalisation and shining-through (Hansen-Schirra, 2011). Nikolaev et al. (2020) found that the relative contribution and prevalence of each tendency depended on the distance between the source and target languages. Translations from structurally-similar languages were found to demonstrate greater conformity to the TL norms and were more predictable, while translations from structurally-divergent SLs contained more non-idiomatic features making them more entropic and unusual in terms of lexical density, mean sentence length, frequencies of conjunctions and passives, etc.

It seems reasonable that, according to some authors, a reliable translationese study should involve an aligned parallel corpus and an interface with findings from contrastive analysis, which is not always the case. Evert and Neumann (2017) state: "it is methodologically impossible to determine differences between translated and non-translated texts without comparing the realization of a feature in the matching source text" (p.50). This stance is reminiscent of an early Toury's opinion that "an exhaustive contrastive description of the languages involved is a precondition for any systematic study of translations" (Toury, 1980, p.29). When the methodological setup of a study complies with this requirement, a difference can be made between adaptation (covert translation, when diver-

ging SL frequencies of a feature are fully adopted to the TL norm) and *over-normalisation*, an exaggeration of target language norms leading to unusual overuse of typical TL patterns (Evert and Neumann, 2017).

Finally, it is not unusual to group trends in translational behaviour into shining-trough and source-language-independent effects (Koppel and Ordan, 2011) or SL/TL-independent translationese (Kunilovskaya and Lapshinova-Koltunski, 2020; Kunilovskaya and Pastor, 2021), i.e. "forms of translationese that cannot be traced back to the respective source language or to individual source texts" (Evert and Neumann, 2017, p.54).

On top of the four major trends defined on page 18 and interference (alternatively called shining-through or transfer as shown above), Chesterman (2004) mentions eight other self-explanatory claims such as *lengthening*, *reduction of repetition*, *'unique items hypothesis'*, *untypical lexical patterning*, etc.

In the current terminological conventions these disparate phenomena are known as *tendencies in translational behaviour* (see the use of the term in Laviosa, 2008; Cappelle and Loock, 2017) that result in a frequency difference between translations and comparable non-translated texts in the target language.

**In this thesis translationese is broadly understood as any statistical deviations of translations from comparable non-translations**, regardless of their origin and nature. The main focus of our investigation is the correlation between the amount of translationese captured by the proposed feature sets and available quality labels. The analysis of translationese trends is viewed as a preliminary step.

The original interpretation of the term *translationese* by Gellerstam (1986) was purely descriptive and as such should be *devoid of any derogatory connotations*. According to him, translationese is not the result of poor translation; it is a statistical phenomenon which consists in the diverging distribution of certain features. We agree that diverging patterns in translations are their inherent inevitable property: it cannot mean poor quality. However, a relative amount of translationese represented as distance from the expected standards (and measured in De Sutter et al. (2017), for example) might have implications for quality. In-depth account of theoretical and empirical foundations for this statement appear in Section 2.2.

In the related literature this term is occasionally used as evaluative. For Baker (1993) translationese describes "cases, when an unusual distribution of features is clearly a result

of the translator's inexperience or lack of competence in the target language" (p.249). Blum-Kulka (1986) distinguish 'true hybrid' (as a result from 'positive' translation solutions) and "inadequate text which exhibits features of translationese, resulting from a lack of competence" (p.167).

Another conflicting use of the term *translationese* is metonymic: in a number of contexts, especially in the publications in computational domains, 'translationese' is used interchangeably with 'translated texts'. Explanations like "translationese can be defined as any text that has been translated into another language and thus bears certain linguistic features such as simplification, normalisation and explicitation" are not uncommon. Nikolaev et al. (2020) clearly uses translationese to refer to translated texts: 'translationese as a special language form', 'predictability of translationese' (see also "... contains elements of both original text and translationese"in Carter and Inkpen (2012, p.52)). In this thesis we use *translationese* to mean translationese properties of translations. Popescu (2011) used *translationese identification* to refer to a binary text classification (translations vs. non-translations in the TL). Throughout this thesis we will adopt the term *translationese classification* or *translation detection* to denote this experimental setup following a recent practice in computational translationese studies (Pylypenko et al., 2021).

### 2.1.2 Features of Translation

In this work, 'features' are defined in line with the computational linguistics convention as quantitative characteristics of samples/observations (documents) used to represent these samples in a machine leaning setup or for statistical analysis. A feature is a numeric evaluation of a particular parameter of a sample. To put it simply, features are names of the (non-metadata) columns in a data table, where rows represent each observation. However, in traditional translation studies the term '(universal) features' can be used in a different meaning and context. For example, Zanettin (2013) employs this term to refer to simplification, explicitation, etc (p.23), while practical frequency-based operationalisations of these hypotheses are called 'formal operators'.

In some descriptions, a feature can refer to a multi-dimensional vector, and not one value. Puurtinen (2003) investigates three "features", each of which is actually associated with frequencies of several or many items: complex non-finite constructions, clause connectives, or keywords. While in Puurtinen (2003) each of thirteen connectives is invest-

igated separately, Xiao (2010), inter alia, compares translated and non-translated corpora using cumulative normalised frequencies of dozens of listed conjunctions in each of 2,000-word text chunks in their study. Thus, in Xiao (2010) setup the same feature (connectives) would be represented by just one number for each sample, and in Puurtinen (2003) it is a multi-dimensional vector. This type of features (esp. frequencies of n-grams, POS or mixed) is commonly referred to as text representations rather than features (Bernardini and Baroni, 2005; Kurokawa, Goutte and Isabelle, 2009).

The choice of features largely depends on the research methodology and research goal. A distinction can be made between corpus-based and machine learning computational paradigms, each with their own approaches to data collection, research design and feature selection. We overview these paradigms below.

**Corpus-based and computational approaches** **Corpus-based studies** rely on manual analysis of (parallel) concordances and statistical comparison of translational and reference corpora. By necessity they use a few well-picked linguistic forms or groups of them.

For example, based on contrastive knowledge about grammatical structures in English and Portuguese, Santos (1995) explores the use of tense and aspect verb forms in clearly defined translationally interesting types of contexts. These contexts inspire specific theoretical expectations of translational behaviour.

A typical case is the overt use of linguistic devices that are expressed (and/or obligatory) in the ST and optional in the TL. Explicitation of TL-optional cohesive markers and changes in the frequencies of specific semantic group of discourse markers has become a very fruitful topic for corpus-based translation studies (Olohan, 2001; Nakamura, 2007; Castagnoli, 2009; Becher, 2011; Kunilovskaya, 2017). These studies demonstrate that cohesive markers are good indicators of translationese, and they have been invariably included into more extended feature sets (Puurtinen, 2003; Rabadán, Labrador and Ramón, 2009; Xiao, 2010; Redelinghuys and Kruger, 2015), including in the ML contexts (Volansky, Ordan and Wintner, 2015). However, some caution is required with regard to this feature when studying translations from TED talks corpus, an easily available, reasonably large parallel corpus spanning many language pairs, specifically adopted to translationese research by Karakanta, Vela and Teich (2018). TED talks corpus con-

tains translations produced for subtitling purposes. The one of the instructions included into the guidelines for volunteer translators, helping them to meet the required length requirement, is to lose connectives[1].

In a similar vein, earlier corpus-based studies have established consistency of such translationese indicators as type-to-token ratio (TTR), lexical density, mean sentence and word length, frequency distributions of most frequent words (known as 'list head' from Laviosa, 1998). These features were tested on various language pair, including English-to-Chinese (Xiao, 2010) and translated English (Kruger and Rooy, 2012; Redelinghuys and Kruger, 2015).

Corpus-based approaches are characterised by acute interest to catching translationese on the lexical level. Apart from lexical profiles and distributions of words from various frequency bands, researchers compare counts of neologisms, loanwords, hapax legomena, cognates, formal vs. informal vocabulary, 'common lexical bundles' (Kruger and van Rooy, 2010), phrasal verbs (Cappelle and Loock, 2017) to study creativity and standardisation in translation.

Statistical methods used to reveal translationese trends include significance tests and effect-size coefficients, analysis of vaiance (ANOVA), measures of log-likelihood (Xiao, 2010) and $\chi$-squared (Cappelle and Loock, 2017), correspondence analysis (Delaere, De Sutter and Plevoets, 2012).

More sophisticated corpus techniques are used by De Sutter et al. (2017) and Evert and Neumann (2017). These studies are focused on visualisation as a research method and make use of unsupervised and mildly-supervised ML approaches, including PCA and Linear Discriminant Analysis (LDA).

In terms of features, they build extended lists of well-motivated features, bringing together the previous findings in the field. Both studies make explicit use of language-specific features.

Evert and Neumann (2017) set out to explore the impact of translation direction on the properties of translations using a register-balanced bidirectional English-German corpus. Their methodology combines corpus-based statistical analysis, factor analysis and text classification.

According to the authors, there are two possible experimental setups to achieve their

[1] https://translations.ted.com/How_to_Compress_Subtitles

goal: it is either access to the SL side of a parallel corpus or availability of translations from many SLs into the same target. They go with the former and make a point of employing cross-linguistically comparable features to be able to compare their frequencies in the SL, TL and in translations. The latter setup is very popular in translationese studies, too, but most of the work is done on the same corpus, EuroParl (Koehn et al., 2005) and its versions (Van Halteren, 2008; Koppel and Ordan, 2011; Carter and Inkpen, 2012; Volansky, 2012; Lembersky, Ordan and Wintner, 2013; Rabinovich, Ordan and Wintner, 2017, to name just a few).

Feature selection in Evert and Neumann (2017) is guided by contrastive analysis to ensure cross-linguistic comparability and by variational linguistics. The authors treat translated registers as language varieties and expect the features to be fit for catching register distinctions in their textual data. Table 2.1 has a few examples from their set, counting 22 features.

| Feature | Description |
|---|---|
| adja_T | attributive adjectives, per no. of tokens<br>English: all tokens tagged "JJ.*" (general adjective)<br>German: all tokens tagged "ADJA" (attributive adjective) |
| finites_S | finite verbs, per no. of sentences<br>all items receiving the tag (chunk_gf="fin") in the manual annotation |
| lexical.TTR | lexical type token ratio<br>all lemmatized items tagged as noun, full verb, adjective and adverb<br>over noun, full verb, adjective and adverb tokens per text |
| modals_V | modal verbs, per no. of verbs<br>all items receiving the tag "VM.*" (modal verb) |
| nominal_T | nominalizations, per no. of tokens<br>English: all tokens tagged "N.*" ending on *-ion, -ism, -ment, -ness*<br>German: all tokens tagged "N.*" and ending on *-ung, -heit, -keit, -ismus* |
| pronouns_T | personal pronouns, per no. of tokens<br>listed items for each language tagged "PP.*"/"PPER" |
| subordination_T | subordinating conjunctions, per no. of tokens<br>English: all tokens tagged "CS.*"<br>German: all tokens tagged "KOU.*" or "KOKOM" |

Table 2.1: Examples of features proposed by Evert and Neumann (2017)

This is the approach to feature selection we generally follow in this project.

The same theoretical stand with regard to feature selection is consistently maintained in a series of publications by Lapshinova-Koltunski and Vela (2015); Lapshinova-Koltunski (2017), who uses a smaller set of simpler features, intersecting with Evert's and linked to components of communicative situation (field, tenor, mode) underlying register variation according to Halliday and Hasan (1989). In particular, her feature set includes (i) a subset of PoS n-grams (nominal and verbal chunks); (ii) nominalisations and general nouns; (iii) semantically categorised modal expressions; (iv) evaluative patterns capturing author's epistemic stance, along with other features such as lexical density and frequency

of pronouns that are typically included in both translationese studies and register analysis.

De Sutter et al. (2017) compares student and professional translations to non-translations using PCA plots and ANOVA analysis. The features in De Sutter et al. (2017) are much less linguistically-motivated, but most of them are easily-extractable and language-independent. The authors admit that the selection is rather arbitrary (p.29). Their features include a lot of surface phenomena such as basic frequency information on different PoS categories (lexico-grammatical features), TTR, lexical density, hapax legomena, Zipf-based general word-frequency measure and the total number of the 100 most frequent 3- and 4-grams, which reflects the degree of syntagmatic patterning. The statistical results demonstrate that only seven out of 30 features capture distinctions between students and professional (in English-to-French scenario).

A focus on student translations is not uncommon in corpus-based studies (Nakamura, 2007; Rabadán, Labrador and Ramón, 2009; Castagnoli, 2009; De Sutter et al., 2017; Kunilovskaya, Morgoun and Pariy, 2018) and computational studies (Ilisei et al., 2010; Lapshinova-Koltunski, 2019; Bizzoni and Lapshinova-Koltunski, 2021). The findings often show that students exhibit more translationese than professionals. We will discuss this aspect of translationese studies in Section 3.1.2.

As shown above, the use of interpretable, theory-based features is typical for corpus-based approaches. Feature selection is often guided by cross-linguistic considerations (as in Santos, 1995; Evert and Neumann, 2017; Cappelle and Loock, 2017) and anticipated "shining-through" effect or by translation-study-related theory of translation universals (Redelinghuys and Kruger, 2015; Redelinghuys, 2016; Delaere, 2015).

**More recent translationese studies moved from corpus-based approaches to computational methods**, and rely on supervised and unsupervised machine learning techniques. Depending on the major research goal, **computational translationese studies** can be designed with the view to shed light on the nature of translational behaviour or as a computational challenge seeking to solve an associated NLP task, e.g. translation detection and especially, SL detection, a task that has developed a lot of traction recently. For example, Sominsky and Wintner (2019) acknowledge: "Our main goal here was to guarantee best performance, even the cost of interpretability" (p.1134).

A good example of a theory-oriented ML-based translationese study is presented in Volansky, Ordan and Wintner (2015). They state: "Our main objective, however,

is not to produce the best-performing classifiers. Rather, it is to understand what the classifiers can reveal about the nature of the differences between O [non-translations – MK] and T [translations – MK]" (p.110). Their work has become influential in subsequent research and is often used as a baseline due to a clear description of the methodology, including extraction procedures for a fairly extended set of features. For each text in their experiment, they extract 38 features, with some being a single value (i.e. TTR) and other being multi-dimensional (e.g. for POS trigrams they extract 300 most frequent sequences). Each feature is linked to one of the four translation universals (simplification, explicitation, normalisation and interference) in a top-down manner to explore the hypothesised trends in translational behaviour. Table 2.2 reproduces their mapping.

It can be seen that the features in Evert and Neumann (2017) are more knowledge-rich than the features in Volansky, Ordan and Wintner (2015). The former implements fairly complex extraction rules to get nominalisations, for example. Naturally, more complex rules raise the issue of extraction procedures reliability. These two studies use their feature sets differently: Volansky, Ordan and Wintner (2015) tests each feature (or feature set) separately, while Evert and Neumann (2017) argues for a multivariate analysis and uses all feature values as a feature vector representing each instance in the experiment.

The second difference between these two studies is the underlying theory. While features in Evert and Neumann (2017) rely on findings in contrastive linguistics and register studies, Volansky, Ordan and Wintner (2015) selects them to represent theoretical categories from translation studies in a top-down manner, by mapping each feature to a translationese trend.

The quest for translationese trends, represented by pre-selected features, has been a very popular design in computational translationese studies (Corpas Pastor et al., 2008; Ilisei et al., 2010; Volansky, Ordan and Wintner, 2015, and most recently (Hu and Kübler, 2021)). These studies returned controversial results. To give some examples, Corpas Pastor et al. (2008) confirmed positive results for some features they associated with simplification, but not for the others. Kruger and van Rooy (2010) reported limited support for the "more explicit, more conservative, and simplified language use in the translation corpus" (p.26), as represented by features selected by them. A recent study by Hu and Kübler (2021) did not find support for either simplification or normalisation.

This is not surprising for at least three major reasons:

| Category | Feature |
|---|---|
| Simplification | TTR |
| | Mean word length |
| | Syllable ratio |
| | Lexical density |
| | Mean word rank |
| | Mean sentence length |
| | N most frequent words |
| Explicitation | Explicit naming |
| | Single naming |
| | Mean multiple naming |
| | Cohesive Markers |
| Normalization | Repetitions |
| | Contractions |
| | Average PMI |
| | Threshold PMI |
| Interference | POS 1-,2-,3-grams |
| | Character 1-,2-,3-grams |
| | Prefixes and suffixes |
| | Contextual function words |
| | Positional token frequency |

Table 2.2: Features associated with each translation trend in Volansky, Ordan and Wintner (2015)

1. the mapping of particular features into descriptive translationese trends can be a matter of debate (as stated in Zanettin (2013, p.25));
2. there can be differences in the extraction procedures;
3. translations from different SLs and in different registers produce diverging translationese patterns.

Therefore, in this research we refrain from assigning individual features (indicators) to

particular trends such as simplification and explicitation in a top-down manner. Instead, we follow a bottom-up approach and identify the indicators of some translationese effects based on the similarity of their frequency pattern in source texts, target texts and reference texts to categorise features as contributing to different translationese trends. Our feature selection and research design is most in line with the third approach to translationese as presented in the beginning of Section 2.1.1 (page 18). We evaluate the features in the context of SL-TL contrast, i.e. as contributing to shining-through or normalisation tendencies.

**Surface, linguistic and linguistically-motivated features** With regard to the amount of knowledge, informing the extraction procedures, the features used in translationese-related research fall into three categories.

**Surface and 'easily extractable' features** do not use any linguistic-theoretical information and/or external resources and are generated from raw text. Most commonly they are character or word n-grams of various order. They can be interpreted at a higher linguistic level and are shown to capture morphological phenomena. Surface features also include mean length of words or sentences as well as TTR.

It is well-known that character- and word n-grams, being lexicalised string features, tend to capture differences in topical domains between translations and non-translations, rather than true distinctions between respective language varieties. This limitation was described in very early computational studies, such as Bernardini and Baroni (2005) and Van Halteren (2008). Although surface features have been shown to deliver nearly perfect classification results (Popescu, 2011), they do not generalise well across domains and capture uninteresting domain distinctions between corpora (Eetemadi and Toutanova, 2015). Researches, interested in theoretical insights, prefer to opt for semi- or fully delexicalised features (Lapshinova-Koltunski and Zampieri, 2018; Nikolaev et al., 2020; Hu and Kübler, 2021). The same practice is adopted in experimental variational linguistics (Diwersy, Evert and Neumann, 2014).

**Linguistic features** are generated using external resources and language models. A very popular approach is PoS n-gram representations (including frequencies of PoS tags, i.e. PoS unigrams) from annotated corpora which have all words replaced with their tags.

In computational research designs, it is not uncommon to rely on simple linguistic

features. For example, Koppel and Ordan (2011) used 300 functional words to achieve staggering classification results on a number of EuroParl language pairs, demonstrating the existence of SL-related "translationese dialects". Rabinovich and Wintner (2015) rely on function words and cohesive markers. Nisioi and Dinu (2013) uses a set of most frequent words (MFW), inspired by computational stylometry and authorship attribution studies.

One of the famous early studies by Baroni and Bernardini (2006) compares the performance of word n-grams and PoS n-grams, and their combinations on high-quality journalistic texts translated into Italian, using linear SVM. They show that the best translationese indicators are distribution of function words and morphosyntactic categories, including personal pronouns and adverbs.

A fairly novel and well-motivated approach to linguistic feature selection relevant to this project is presented in Hu, Li and Kübler (2018). They use delexicalised syntactic features – constituent parse trees and dependency triples – as features. Namely, they use frequencies of (i) each grammar rule extracted from the parse trees, (ii) 'depTriple', i.e. PoS of a head and its dependent along with the dependency relation, and (iii) dependency labels. They found that combined feature sets work significantly better than one feature set alone, but translations could be reliably distinguished from non-translations on only five structural features. Their analysis of top-ranking features proves that syntactic representations yield themselves for interpretations better than PoS n-grams that were interpreted on the syntactic level in previous research.

Recent research makes use of language modelling approaches. An important distinction to be made here is between feature-based and end-to-end feature-learning approaches. The main use-case scenarios are as follows.

First, vector representations from pre-trained models can be used as input features in their own right (e.g. Pylypenko et al. (2021) uses *fastText*[2] word vector models in an SVM classifier) or to initialise a neural network architecture setup to solve the task at hand. For example, Sominsky and Wintner (2019) used a bi-directional Long Short-Term Memory (BiLSTM) seeded with GloVe word embedding vectors [3] for each sentence pair to solve a challenging task of sentence-level translation direction detection. It was shown to outperformed all feature-engineering approaches attempted in their research. A

[2] fasttext.cc/docs/en/pretrained-vectors.html

[3] nlp.stanford.edu/projects/glove/

recent work by Pylypenko et al. (2021) explores the performance of pre-trained neural word embeddings and end-to-end neural architectures (feature-learning approach) in a translationese classification task to show that they outperform hand-crafted features but capture different aspects of translated language. Based on their multilingual experiments, they conclude that while (deep) feature-learning approaches provide evidence for the existence of universal translationese specificity, hand-crafted features used in the experiment can hardly explain it. Interestingly, their *hand-crafted* features include crude surface features (average word length) and features from other linguistic representation (unigram bag-of-PoS, n-gram frequency distribution).

Second, language models can be used to generate information-theory-related 'surprisal' features (language modelling features) such as log-probabilities and perplexities (Rubino, Lapshinova-Koltunski and Van Genabith, 2016; Bizzoni and Lapshinova-Koltunski, 2021; Pylypenko et al., 2021) or entropies (Nikolaev et al., 2020; Hu and Kübler, 2021). This approach builds on the assumption that if translated texts are different from non-translations they should be more difficult to predict for a language model trained on the originally-authored texts in the TL. The language models can be trained on delexicalised corpora, i.e. on sequences of PoS or syntactic tags. It allows (i) to minimise domain-related differences and (ii) to reduce the model vocabulary making it possible to learn patterns on relatively small training corpora. Rubino, Lapshinova-Koltunski and Van Genabith (2016) computed surprisal, complexity, and distortion features based on flattened constituent parse trees. In Bizzoni and Lapshinova-Koltunski (2021), language model (LM) perplexities were interpreted as a measure of translationese, but the results were controversial: the authors reported consistently higher perplexity on professional translations than on student translations across a number of registers. It means that their neural model trained on non-translations in the TL found professional translations more unusual than student translations.

Third, Chowdhury et al. (2020); Chowdhury, Espãna-Bonet and van Genabith (2021) solve a computational typology task of reconstructing phylogenetic language trees from translations, first attempted by Rabinovich, Ordan and Wintner (2017). Following the approach in Bjerva et al. (2019), they generated embeddings for the whole language varieties (or, in Koppel and Ordan's terminology, translationese dialects (Koppel and Ordan, 2011), i.e. translations from many languages into one language) and measured

distances between these embedding spaces. The models were trained on delixicalised versions of originally-authored English and English translated from a dozen of languages to demonstrate that translations retained enough traces of respective SLs to reproduce linguistically-informed language typology. As expected, translations from SLs belonging to the same language family clustered together, and the vector spaces for structurally similar SL and TL had more isomorphism.

**Theoretically-motivated or linguistically-motivated features** usually stand to mean features which are designed to capture specific linguistic phenomena. They are usually backed by findings from contrastive linguistics or from translation studies, and require relatively complex extraction rules and external resources, including pre-defined lists of items and additional corpus annotation. Many of these features were presented in the paragraphs on corpus-based approaches in translation studies (pages 23-30).

In this research we use a set of explicit linguistically-motivated and interpretable features that are tested to capture translationese in student and professional translations in a quality estimation task. To demonstrate their competitive performance, we compare them with term frequency–inverse document frequency (tf-idf) weighted word n-grams, on the one hand, and with representations from pre-trained embedding models, on the other, for both tasks. The description of our selection of features appears in Section 2.3.

## 2.2 Translationese and Quality

This section presents the rationale behind the hypothesis put forward in this research. What theoretical and empirical evidence is there to assume that translationese features might also be predictive of translation quality?

The link between translationese and translation quality has been assumed in corpus-based translation studies ever since translationese has become an active research field. It is typical to use the term *translationese* in derogatory contexts, similar to other words built after this derivational pattern: *legalese, journalese, academese, post-editese*. All these nominations imply hyperbolic, unnecessarily exaggerated use of features characteristic to the respective communicative situation, often opposed to plain language: legal matters, news, academic writing, etc.

Although translations are always different from non-translations bearing inevitable

marks of the translation process, the magnitude of these differences can be related to quality.

**Evidence from previous studies** There is a number of corpus-based studies that correlate the individual translationese trends and quality scores/translation expertise[4].

In particular, simplification and explicitation were tested as possible indicators of translation quality with a view to improving teaching methods and assessment criteria at postgraduate level. Simplification was found to correlate with lower-scoring translations and explicitation was found to correlate with higher-scoring translations (Scarpa, 2006).

Rabadán, Labrador and Ramón (2009) interpreted translationese departures from TL norms as quality-related. They identified three grammatical translationese-prone areas in English-to-Spanish translation (quantifiers, modifiers of nouns and the translation of the English Simple Past form) and demonstrated how to use corpora to corroborate disparities between translated language and non-translations. Interestingly, they called for an implementation of language-pair-specific statements as Translation Quality Assessment (TQA) tools that could contribute to the systematisation and objectivisation of translation assessment. Consider the examples of the proposed conditional statements for English-to-Spanish translation:

1. The lower the number of formal options chosen from those available in Spanish to translate intensified quantification, the higher the degree of simplification and the less accurate the translation, and vice versa;
2. The lower the number of *de*-phrases, the higher the number of pre-modifying adjectives in translated Spanish, the higher the degree of interference and the less idiomatic the translation, and vice versa;
3. The smaller the disparity between native and translated usage in the use of particular grammatical structures associated with specific meanings, the higher the translation rates for quality.

Redelinghuys and Kruger (2015); Redelinghuys (2016) explored the differences between professional translators and students with regard to hypothesised trends in translational

[4]Note that some researchers exercise caution in using professional varieties of translation (students and professionals) as a proxy for quality

behaviour (explicitation, simplification and normalisation, levelling-out). The authors avoided referring to levels of expertise as quality categories. Instead they talked about using "features of translationese as indicators of translation expertise"(Redelinghuys, 2016, p.192).

Unlike the previous work, De Sutter et al. (2017) treated student and professional translations as quality-related categories. The instances in their corpus-based study were represented as feature vectors valued on typical translationese indicators from previous research. They visually explored locations of individual translations relative to professional standards (professional translations and comparable non-translations) in a low-dimensional space and performed univariate analysis to produce aggregate judgments about deviations in student translations. However, no predictive model was attempted. They also reported conflicting results on their two case-studies attributing the confusion to feature selection and calling for less shallow, more linguistically-motivated approach. They highlighted that translationese-based approach to quality reflected only the fluency aspect of quality.

At the onset of machine learning approach to translation detection, Baroni and Bernardini (2006) briefly mentioned developing an automatic translationese spotter to be used in translator education.

In a ML study based on a register-balanced English-German corpora of human and machine translation, Lapshinova-Koltunski and Vela (2015) found that human professional translations might deviate from the conventions of non-translated TL registers, and therefore could not be relied upon as the only gold standard for MT quality evaluation: TL comparable texts belonging to the same registers should also be considered.

The most straightforward application of translationese indicators to translation quality prediction task is reported by Aharoni, Koppel and Goldberg (2014) in the field of MT. They proposed a generic machine translation quality estimation technique which used the accuracy metric from a translation prediction task. The result of the study indicated that translation detection accuracy was strongly correlated with the evaluations scores, yielding $R^2 = 0.774$. The ratio of sentence pairs identically ordered by the proposed metric and human evaluators (based on French-English data from WMT-2013) was 0.846. Quite unexpectedly, they used binary features, denoting the presence or absence of PoS n-grams and each of 467 function words.

The intimate link between translationese and quality is manifested in a computational study that solves the reverse task of predicting translations using MTQE features. In other words, the authors used indicators of translation quality to predict student and professional translations (Rubino, Lapshinova-Koltunski and Van Genabith, 2016).

**Fluency as overall quality?** Generally, it seems reasonable to contend that the smaller the disparity between originally-authored and translated language, the higher the translation quality.

It is obvious that translationese-related defects reflect only one aspect of quality, namely, fluency. The analysis of pragmatic and semantic relations between ST and TT is not involved in translationese-related experiments rendering acceptability and accuracy inaccessible for translationese-based measures.

It can be argued that fluency is major quality factor in human perception. Grammatical disfluencies and raptures to the natural text flow, which sometimes turn translations meaningless, immediately catch an eye and have an impact on the overall perception. Even if translationese does not result in ungrammatical structures, it generates strange and unnatural linguistic forms that are recognised as literal translations by professional bilingual and general public. 'Does not sound right' effect in translations can be linked to the concept of non-binary translation errors put forward by Pym (1992). Repeated manifestations of translationese can be considered a distributed translation error, which can only be diagnosed based on document-level statistical analysis like the one proposed by De Sutter et al. (2017).

The strong preference for fluency in translation assessment can be explained by the current dominance of translator's invisibility as a conventional norm in translation as opposed to a more marginal strategy of 'foreignisation' (Venuti, 1995). Although it has been shown that professional tolerance to interference might vary across registers in one language and across language pairs (language prestige effect), the general expectation for a high-quality professional translation is to be covert and as much blended with the TL norm as possible.

At the same time, it is obvious that semantic relations between ST and TT cannot be entirely ignored in translation quality estimation. Accuracy might be a more important aspect of quality that fluency: it is better to have a poorly-written translation that

accurately conveys the content of the ST than a very fluent translation that misinforms the reader about the meaning of the source. However, when it comes to evaluation experiments, accuracy-fluency distinction might be more theoretical than practical. Human annotators of MT produced highly correlated scores for accuracy and fluency (Callison-Burch et al., 2007), and there is a trend to dispense with this distinction in the annotation setups acknowledged in a recent survey of the MT evaluation metrics (Chatzikoumi, 2020). Also, this problem is recognised in Daems et al. (2017): their solution was to keep the annotation of the two aspects of quality separate.

Besides, it is difficult to imagine a human translation produced in good faith that would completely and seamlessly misrepresents the content of the source. This last argument can be supported by the results obtained in the studies on student translations. Carl and Buch-Kromann (2010) found that the major difference between learner and professional translations were about the degree of fluency. De Sutter et al. (2017) quotes Loock (2016) as concluding: "some deviant linguistic characteristics of translations written by advanced students might be symptomatic of the overall quality of their translations" (p.28).

The error statistics in the RusLTC[5], used as a source of student translations in this thesis, indicates that only 23.7% of errors (less than a fourth) were content (or accuracy errors), while the majority of annotated translation solutions had to do with style, cohesion, lexis or sentence structure, i.e. with fluency.

**Relevance of translationese for MT** The impact of translationese on the outcomes of MT has been demonstrated with regard to the training data, and with regard to the quality of test sets. A series of papers have demonstrated that it was important to take into account the directionality of translation when training MT systems[6] (Kurokawa, Goutte and Isabelle, 2009; Lembersky, Ordan and Wintner, 2012, 2013; Stymne, 2017).

Translational deviations from the target language were shown to have an impact on the reliability of MT evaluation. The systems demonstrated better results when test sets included reversed parallel sentences (Zhang and Toral, 2019; Graham, Haddow and Koehn, 2020). Translated documents in the test sets proved less challenging for MT engines.

[5] www.rus-ltc.org/static/html/about.html

[6] using all available language pair data regardless the translation direction is typical to increase the size of the training set

## 2.3 Proposed Features and Representations

### 2.3.1 Feature Selection

In choosing features to test our hypothesis, we opted for **indicators of translationese that are at the same time 'usual suspects' for human translation fluency errors** like it was done by Rabadán, Labrador and Ramón (2009).

Practical textbooks for translation students typically address translationese-prone areas hypothesised from cross-linguistic contrasts, including contrasts in the frequency of shared features, and by observations of typical suboptimal solutions by inexperienced translators. An analysis of a representative sample of 46 translation textbooks described in Kunilovskaya et al. (2022, see Table 4) demonstrated that most of the anticipated translation problems were grammatical phenomena (25 issues in 39 books). Lexical and stylistic issues attracted less attention in the examined publications (16 issues in 31 books). However, the analysis of error-annotated student translations revealed higher salience of lexical problems, as well as a different ranking for more frequent grammar difficulties.

Taking into account the focus in the textbooks corrected with the analysis of actual errors observed in real-life student translations in our corpus, as well as considerations from the previous corpus-based and computational research on translationese, described above, we propose a set of 60 morphosyntactic features (Section 2.3.2) and two sets of abstract lexical features: 24 collocational features (Section 2.3.3) and 10 features from n-gram lists and n-gram language models (Section 2.3.4).

Various components and versions of these feature sets were used to analyse English-to-Russian and English-to-German translation in four registers (Kunilovskaya and Lapshinova-Koltunski, 2020), to explore its robustness to register variation in English-to-Russian professional translation (Kunilovskaya and Pastor, 2021) and to solve a SL detection task based on literary translations into Russian from 11 languages (Kunilovskaya, Lapshinova-Koltunski and Mitkov, 2021*a*).

This section starts with a few examples of translationese-related problems observed in student translations.

Typical translationese-prone areas for English-to-Russian translation include such morphosyntactic categories as the overuse of relative clauses, copula verbs, modal predic-

ates, analytical passives, generic nouns, all types of pronouns, etc. Consider the following examples:

(2.3.1) Necklaces, at first as pectorals that covered the whole chest, evolved from the prehistoric pendants.

Ожерелье – первое нагрудное украшение, *которое* занимало место на всей груди, *которое* стало основой для подвесок

'Necklace – first chest decoration, *which* covered the whole chest, *which* became the basis for pendants.'

(2.3.2) ... there are many self-employed people who manage to get money from others by means of falsely pretending to provide them with some benefit or service ...

Более того, *есть* много *людей*, работающих на себя, *которые* получают деньги обманным путем ...

'Moreover, many *people* are, working for themselves, *who* get the money in a deceitful ... way'.

(2.3.3) ... differences in self-efficacy may simply mean that some teachers struggle to identify solutions to problems beyond their circle of control.

... разница в самооценке *может* означать лишь *то, что* некоторые учителя испытывают сложности в нахождении решений задач за пределами *того, чем* они *могут* управлять.

'... difference in self-evaluation *can* mean only *that that* some teachers run into difficulties in finding solutions to tasks beyond the scope of *that what* they *can* control'.

(2.3.4) The results are also of interest from a contrastive linguistics perspective, providing multivariate evidence that the difference between two languages is not only observable in features that only exist in one language but also emerges from the distributional patterns of comparable features. It was difficult and exhausting to see.

*Это было* тяжело и утомляюще пытаться видеть.

'*It was* hard and exhausting to try to see'.

These examples from RusLTC demonstrate a number of translation solutions that explain the increase in the frequency of TL items that are less frequent in non-translated TL than their literal counterparts in the SL.

Probably, none of the translations in the examples can be considered ungrammatical in Russian, but there is a clear Master Yoda-style foreign sound to them. Beware that the back translations may come across as perfectly acceptable sentences, because the translations are very literal in the first place.

In Example 2.3.2 a generic noun (*people*) is rendered with a less frequent literal *люди*, instead of a possible structure with zero subject or other more acceptable ways of expressing unspecified subjects.

English and Russian have contrastive ways of expressing subjective modality: modal verbs are a less common choice in non-translated Russian, which prefers parenthetical epistemic markers. The translation solution in Example 2.3.3 carries over a typical English modal predicate.

Example 2.3.4 has notorious literal renderings of the structures with introductory *it* and copula verbs in contexts, where a zero copula is typical in Russian. This contributes to a boost in frequencies of pronouns and copula verbs in translated Russian. Besides, such renditions have a strange word order, which usually interferes with the smooth flow of information in the text.

Another source of surplus function words, including pronouns, is the tendency to unpack the information from various concise English structures using strings of relative clauses, instead of repackaging the information in a more natural way (see Examples 2.3.1 and 2.3.2).

Below, we discuss principles and criteria that guided feature selection in this work, followed by a description of extraction procedures.

Special effort was made to keep the features **cross-linguistically comparable** as suggested by Evert and Neumann (2017). While implementing extraction rules for a number of features we had to take into account diverging annotation rules and possible variety of expression for a given category in either English or Russian. For example, English passive verb form corresponds to at least four Russian lexico-grammatical structures.

The rationale behind this decision is an attempt to reveal the most notorious effect in translation, namely, 'shining-through', a translational tendency to reproduce SL patterns

and frequencies rather than follow the TL conventions. This form of translationese can be established by comparing the distributions of feature values between English source texts, translations into Russian and comparable Russian non-translations. Besides, this allows to put the three corpora into a shared feature space to explore the distances between them visually.

Cross-linguistic analysis yields interesting contrastive findings. Languages differ not only in what they have to express or do not express, but, more importantly from translational perspective, how they distribute comparable isomorphic features. For example, a study on semantic groups of connectives revealed diverging preferences in English and Russian with regard to marking relations between discourse units. In previous research, we showed that English had a tendency to mark contrast (*however, but, nonetheless*)more often than Russian, while Russian had more explicit markers of inference (*therefore, because, that is why*) (Kunilovskaya, 2017).

As it was shown above, we view translations as a TL subsystem, and, therefore, we explored approaches to computational representations used in variational linguistics to find a broad intersection between their features and features from translationese studies. In particular, we considered features and extraction procedures suggested by Biber (1995); Nini (2015) (for English) and Katinskaya and Sharoff (2015) (for Russian). Note that we had to limit ourselves to the phenomena that are **reliably extractable**, among other considerations. While the extraction of abstract lexical features is straightforward, the rules for morphosyntactic features were manually double-checked on our corpora to ensure that they return the expected results. The accuracy of feature extraction is largely dependent on the accuracy of the automatic annotation. Care has been taken to filter out noise by using empirically-motivated sets of function words; recurrent annotation errors were taken into account where possible. More details on annotation and extraction are given in Section 2.3.2.

We also took into account the general requirements to meaningful investigation of translationese, formulated by Volansky, Ordan and Wintner (2015) and reinforced in subsequent research: features should reflect **frequent** linguistic characteristics, be **content-independent** (delexicalised) and be **interpretable**. Our purpose is to understand translational behaviour better by verifying existing intuitions and possibly uncovering new trends in student translations affecting perceived translation quality.

Finally, an attempt was made to reduce collinearity of morphosyntactic features and avoid features that go together in most cases. For example, coordinating conjunction (*cc*) is a syntactic relation that is assigned to items with the PoS tag *CCONJ*; either *cc* or *CCONJ* should be extracted. The syntactic relation named *conjunct (conj)* might be overlapping with *cc* to an extent, but it also accounts for elements coordinated with a punctuation mark (a comma). At the same time we included features that might lack a theoretical or empirical support at the moment, hoping for new insights to be obtained in a bottom-up matter.

Each text instance in our corpus is represented as a sparse feature vector with the maximum dimensionality of 94. To explore the competitiveness of these features, we compare their performance with other types of frequency-based features – *tf-idf* representations and the document-level features from *QuEst++* – and with embeddings from pre-trained neural models.

### 2.3.2 Universal Dependencies-based Features

The first subset includes 60 features extracted from Universal Dependencies (UD) annotations (UD features). They reflect morphological and syntactic structure of language. The Universal Dependencies framework (Straka and Straková, 2017) is selected for annotation in this project because it offers consistent annotation of similar constructions across languages and uses universal tags for structural categories based on typological studies that allow to bring out "cross-linguistic parallelism across languages and language families"[7]. Further details on annotation and preprocessing are given in Section 3.2.4.

The values for UD features are either normalised frequencies of various UD tags and their combinations or measures/ratios and means derived from syntactic trees.

In this section we offer two views of the UD-based feature set:

1. with regard to the type of linguistic phenomena each feature represents (see Table 2.3), and
2. with regard to the amount of engineering effort and manual intervention employed to obtain feature values.

[7] universaldependencies.org/introduction.html

Appendix A has a full alphabetic listing of features by their shorthand names (codes) and more details on the extraction for each feature. Where possible, we provide references to relevant previous work apart from what was discussed above, to explain why specific features were included in the feature set.

Types of general linguistic categories reflected by the proposed morphosyntactic features are presented in Table 2.3.

| type | number | list of features |
|---|---|---|
| word forms | 8 | past tense, passive voice form, finite and two non-finite forms of verb (infinitive and all participles), deverbal nouns, superlative and comparative degrees of comparison |
| word classes | 9 | nouns, personal, indefinite/total, possessive and demonstrative pronouns, adverbial quantifiers, coordinate and subordinate conjunctions, adpositions |
| discourse markers | 5 | contrastive, additive, causative-consecutive, temporal-sequential connectives and epistemic stance markers |
| types of clauses | 8 | adjectival clauses and relative clauses as a separate feature, adverbial clauses, clausal complements with or without own subjects (ccomp and xcomp, respectively), asyndetically joined elements in a sentence, negative and interrogative sentences |
| dependencies | 22 | adjective/participles and other words in attributive function (attrib and amod, respectively), modal predicates, auxiliary verbs, passive voice auxiliary, copula verbs, appositional modifier, etc. |
| text measures | 8 | lexical type-to-token ratio and lexical density (based on disambiguated content types), mean hierarchical and mean dependency distances, number of simple sentences, number of clauses per sentence, sentence and word length |
| TOTAL | 60 | |

Table 2.3: Types of linguistic information by language level captured with UD features

We extract *eight grammatical and lexico-grammatical forms*. They include four fea-

tures designed to capture degrees of nominalisation (finite form > participle > infinitive > deverbal noun). The respective 'deverbalisation scale' was suggested by (Korzen and Gylling, 2011), who explored cross-linguistic differences between Italian and Danish, and their implications for translational behaviour. Similar differences in information packaging are discussed in connection with sentence splitting in language pairs with German and English or Norwegian (Fabricius-Hansen, 1999). In English-Russian language pair there are two prominent translational tendencies in this respect. Russian is known to have a stronger preference for less verbal structures in informational and more formal registers. This feature is often overemphasised (over-normalised) in translation leading to unnatural use of deverbals, often ridiculed in the translation textbooks (e.g. *Вычерчивание им линии происходит с усердием ученика, высунувшего язык от натуги.* [Drawing of a line by him happens with the diligence of a student sticking out tongue from the effort.]) This might lead to the increased frequencies of deverbal nouns in light verb constructions (compared to Russian non-translations). On the other hand, the tendency to copy English syntactic structures (interference) results in carrying over sentences with finite clauses where a nominalisation transformation can be expected. Due to this, we can expect higher frequencies of finite verbs and more clauses per sentence. For example, it can be predicted that translation students would keep the adverbial clause in *The first theme, on how data and theory interact, interested me the most.* to get something like *Первая тема о том, как данные и теория взаимодействуют, заинтересовали меня больше всего.* [First topic about that, how data and theory interact, interested me most of-all] instead of an arguably more *Первая тема, взаимодействие фактов и теории, интересовала меня больше* [First topic, interaction of facts and theory, interested me most]. The necessity to account for free translator's choice explains why we exclude cases of obligatory use of infinitive forms after in modal predicates from the counts.

The frequency of PoS unigrams, especially for *functional word classes*, is a known and strong translationese indicator confirmed in a number of studies. Our focus is on four types of pronouns and on two types of conjunctions. The rationale for English-Russian language pair includes the following factors:

1. Russian is a pro-drop language, while English is not. Personal pronouns are more frequent in English, especially in the descriptions of human poses and actions (e.g.

*she smiled through her tears*, or *every member of our modern society*), where they are not usual in Russian. Hence, we expect more personal pronouns in English-to-Russian translations.

2. Diverging use of possessive pronouns and pronominalised determiners (e.g. this/that, any, every) is a common problem for novice and amateur translators. It is rooted in the drastic differences between English and Russian in how they ensure the flow of information in the text ('communicative dynamism' (Firbas, 1992)) and signal the distinction between topic and comment (and/or theme and rheme). While English relies on the category of determination to a great extent, Russian mostly employs word order as means to convey the functional sentence perspective. This issue invariably attracts attention of translation scholars: Mona Baker has a chapter on in her practical translation textbook (Baker, 2011).
3. The expected increased number of conjunctions can be viewed as collateral to the expected higher frequency of clauses per sentence and increased sentence lengths. Inexperienced translators are known to refrain from sentence splitting; they tend to 'un-package' information from condensed English structures in a very linear manner instead of re-arranging it in a more TL-agreeable way.

For the purposes of this project we define *discourse markers (DM)* as a broad heterogeneous category that includes elements that convey additional meaning to the content of a given discourse unit (a clause, a sentence, a paragraph). According to Fraser (2006) they are "part of a discourse segment but are not part of the propositional content of the message conveyed, and they do not contribute to the meaning of the proposition per se" (p.189). They fall into two major types: connectives, i.e. items that express semantic relations between discourse units and pragmatic markers conveying speaker's attitudes or arranging speaker-hearer interaction.

Connectives are usually classified (with possible variation in terminology) according to type of relation into additive (e.g. *in particular, such as*), causative-consecutive (e.g. *as a result, after all*), adversative (e.g.*instead, despite*) and temporal-sequential (e.g. *eventually, meanwhile*). Pragmatic markers can reflect a much broader scope of pragmatic meanings, but most authors identify the author's epistemic stance markers (reflecting the speakers certainty of the truth value of the proposition) as the most important group (see

Fraser, 2006; Halliday and Hasan, 1989; Biber et al., 2007). In translationese studies based on variational linguistics they are referred to as 'evaluative patterns' (Lapshinova-Koltunski, 2017) and include parenthetical adverbs or phrases like *perhaps, of course* and *на мой взгляд, якобы, полагаю*.

As has been shown above, discourse markers, especially connectives, are a attractive research object in translationese studies: they are content-independent and structurally optional items. Translators are free to choose between explicit and implicit renditions, and their lexical choices are indicative of the adopted global textual and translational strategy.

This study explores five categories of DMs, each represented by their cumulative frequencies extracted by matching elements from pre-defined lists, counting 314 items for English and 269 items for Russian. More examples of each semantic group and their sizes are given in Appendix B.

In this part of feature design we rely on our previous research (Kunilovskaya, 2017), which reports contrasts between English and Russian as to the preferences in explicit marking of semantic relations between discourse units. Based on those findings, as well as previous translationese research, we can expect higher frequencies of all discourse markers, especially adversative and epistemic markers.

The lists were initially produced independently from grammar reference books, dictionaries of function words and relevant research papers. For English we relied on lists in Biber et al. (2007); Fraser (2006); Liu (2008); for Russian: in Shvedova (1980); Priyatkina et al. (2001); Novikova (2008) to name just a few sources for each language. After the initial selection, the lists were verified for comparability. Orthographic and punctuation constraints were used for at least some disambiguation. The output of the extraction procedure was manually checked to exclude greedy matching.

Almost half of all features in this feature set capture various types of syntactic phenomena. We present types of clauses and sentences as a separate subgroup for convenience only. Most of syntactic features are extracted using default UD dependency tags. Recent translationese studies demonstrate that syntactic features might be more informative than morphology in translation detection (Hu and Kübler, 2021) and in SL identification task (Chowdhury et al., 2020).

Finally, we included measures of text complexity, typical in translationese research

such as TTR, lexical density, two measures of parse tree depth, number of clauses per sentence, sentence length, etc.

The normalisation basis used to produce mean values across all sentences in a document varies depending on the type of phenomena. Following Evert and Neumann (2017), raw counts are normalised as follows:

- all morphological categories (word classes) per total tokens in the document (9 features: nouns, four groups of pronoun, two groups of conjunctions, adpositions, adverbial quantifiers);
- all syntactic functions and sentence-level phenomena and measures of complexity per number of sentences (43 features: five groups of DMs, eight types of clauses/sentences, eight text measures, 22 dependencies);
- all verbal forms per number of verbs (6 features: passive, past tense, deverbals, infinitives, participles, finite forms);
- all degrees of comparison per total number of adjectives and adverbs

We used various approaches to extraction, depending on the nature of targeted phenomena and information available from UD annotation for each language.

1. frequencies of all items on pre-defined lists, allowing for some variation in surface form, and disambiguated using position in the sentence and punctuation (6 features), including cumulative frequencies of each semantic group of connectives (*additive, adversative, causative, temposal/sequencial*), *discourse markers of epistemic stance and adverbial quantifiers*;
2. frequencies for combinations of PoS and dependency tags (3 features), including *ratio of nouns and proper names as verb arguments, adjectives/participles in attributive function* and *pronominal determiners*;
3. frequencies of PoS tags, morphological features and/or syntactic tags (and their combinations) filtered on closed lists to filter out noise (8 features), including *adpositions, coordinaive conjunctions, subordinative conjunctions, negative particles or pronouns, possessive pronouns, personal pronouns, indefinite and total pronouns*;

4. frequencies extracted by more complex rules, which are created to enhance cross-linguistic comparability based on the analysis of parallel concordances. They take into account diverging annotation conventions in two languages. These rules extract 13 features:
    - comparative and superlative degrees of comparison,
    - copula verbs (counts for *there + be* were excluded),
    - interrogative sentences (sentences ending in *?!* were included),
    - indefinite and total pronouns (regular expressions are used to extract Russian pronouns; no lexico-grammatic annotation is available for these types of pronouns for either English of Russian unlike other languages in the UD project);
    - deverbal nouns (empirically-motivated negative and positive filters were used);
    - passives (custom-made rules included Russian semantic passive constructions, such as *Меня здесь уважают.* [I-accusative here they-respect.] *I am respected here.*)
    - participles and participial constructions, which have diverging annotation rules in English and Russian
    - infinitives: English rules for 'bare' infinitive were taken into consideration: occurences after model verbs were excluded;
    - modal predicates, which required rules and support list to account for cross-linguistic differences in expressing modality in the predicate;
    - cumulative frequency of clauses per sentence;
    - frequency of simple sentences;
5. ratios based on counts of specified tags, including lexical density, lexical TTR;
6. means for sentence length and word length across all sentences in the document;
7. features measuring syntactic tree complexity (2 features): mean hierarchical distance and mean dependency distance;
8. frequencies of a single morphological attribute or PoS tag (3 features): finite verb forms, past-tense verb form, all nouns
9. frequencies of 24 dependency tags out of 40 available in the UD annotation (including language specific variants available for both languages in this project) scheme[8]:

[8] universaldependencies.org/u/dep/index.html

- six dependencies were dropped because they duplicated the custom extractors or were considered less informative (*cop, cc, csubj, root, det, punct*);
- ten dependencies were excluded because that returned monotone values (all zeroes) for some subcorpora or were very rare (*reparandum, vocative, dislocated, clf, goeswith, dep, expl, list, expl, orphan*)

### 2.3.3 Collocations

Over 30% of the translation errors annotated in student translations fall into the lexical category (see details on the error typology and annotation process in Section 4.4.2). Over half of the issues types revealed through manual analysis of student translations are caused by lexical and stylistic phenomena (collocational differences, lexical ambiguity, translator's false friends or cognate lexemes, differences in conventionalised referencing, cliché and idioms) (Kunilovskaya et al., 2022, Table 5, p.22). In absolute counts, inability to recognise and reconcile cross-linguistic collocational difference is the leading source of disfluency in translations.

To operationalise the specificity of student translations on the lexical level, we introduced two additional feature sets avoiding surface string features and building on the best practices in the field.

The first feature set is designed to reflect collocational properties of translated texts. It includes proportions of n-grams with various degrees of semantic association between components to all types of linear combinations in a document, i.e. to all n-grams.

Our approach to the design of collocational features is inspired by findings in the learner corpus research. In particular, a series of publications by Bestgen and Granger (Bestgen and Granger, 2014; Bestgen, 2017; Granger and Bestgen, 2017) focuses on the use of two association metrics, mutual information (MI) and t-score, to study phraseological development in foreign language learners. They correlated the association scores with the learners' text quality. Their publications introduced the concept of *collgrams* defined as as n-grams with an association score above an arbitrary threshold (Bestgen and Granger, 2014). The authors demonstrated that various categories of association scores had different correlation with the text quality and proficiency levels. The study confirmed their expectation that higher average MI score was a reliable sign of more advanced lan-

guage skills, while t-score was less predictive of text quality. There is some evidence that negatively-collocated bigrams (anti-collocations), i.e. bigrams with the association score below zero, can be an even more reliable indicator than collgrams with a score above a reasonably high threshold, or than all collgrams with score $> 0$.

A separate indicator in their study is the frequency of n-grams that were not assigned a association score, because some of the components were absent in the reference corpus. The relation of this feature to quality is disputable, because unseen tokens can be either errors or instances of creative writing.

Based on their findings, we expect that translations would have a higher ratio of t-score-based highly-associated bigrams, and lower ratio of MI-based highly-associated bigrams than in the comparable subcorpus of non-translations.

This hypothesis is based on the known properties of translation to prefer high-frequency, more established items, known as 'lexical teddy-bears' (Johansson, 2008, p.123), and on the known properties of the association measures. MI is known to favour sequences made of low-frequency items, while t-score assigns higher scores to high-frequency items (Gries, 2010). The observation that collocations of high-frequency words are less typical of native writers should be applicable not only to learner language, but to translations as well.

The ratio of negatively associated bigrams and the ratio of bigrams not seen in the reference is aimed to capture less usual sequences which can be a sign of shining-though or errors, including acceptability in the register (e.g. *тяжелая критика, крепкая основа*).

We also included a collocational feature proposed by Volansky, Ordan and Wintner (2011), i.e. average mean association score for all identified collocations, as well as standard deviation for the association scores to reflect homogeneity in the use of collocations. In our implementation, we averaged association scores for items with a score above zero, excluding anti-collocations from the counts to avoid negative and positive scores canceling each other. A higher average association score indicates more collocations. Based on the previous findings, we expect translations to be less collocated than non-translations, contrary to normalisation hypothesis.

All collocational features were counted for bigrams and trigrams.

To sum up, the collocational features for each association measure and each size of n-gram include:

1. ratio of highly-associated collgrams to all n-grams in the document (the cut-off for high association was set to recommended normalised pointwise mutual information (NPMI) $> 0.5$ and t-score $> 6$),
2. ratio of negatively-associated collgrams (anti-collocations, Evert, 2009) to all n-grams,
3. ratio of all detected collgrams with the score $> 0$ to the total word count,
4. ratio of bigrams absent in the model to all n-grams,
5. average association score for all detected collgrams with the association score $> 0$,
6. standard deviation for the association scores in each document.

In computing the ratios, we relied on n-gram types, not tokens.

This gives us 24 features for each document in the corpus (six features by two scores by two n-gram sizes).

**Extraction details** For each language, each association measure and each n-gram size, we trained a collocation detection model on large register-comparable corpus resources obtained for this purpose and described in Section 3.2.3.

All lexical features were produced from lemmatised corpora. Further on, to reduce the impact of topical differences and of proper names, we removed all identifiers of personal, temporal or spacial deixis, i.e. words referring to a specific time, place, or person in context from both training corpus resources and research corpus.

In practical terms, all words tagged as proper names and their sequences were replaced with a placeholder *PROPN* and all numbers were represented as XXX (e.g. *Борис Николаевич Юрьев* -> PROPN, 1984 -> XXXX, 10,527 -> XX,XXX, *1980-й* -> *XXXX-й*).

Besides, we deleted all punctuation, except end-of-sentence marks.

The models were trained using the *Phrases* module from the *Gensim library*, version 4.2 (Řehůřek and Sojka, 2010), a free open-source Python library designed to process text for machine learning.

The two association measures were implemented as follows.

We used the normalised variant of pointwise mutual information (PMI), suggested by Bouma (2009), which imposed a fixed upper bound on the standard measure of association between components of an n-gram (PMI). Generally, PMI quantifies the "mutual

expectancy" ((Firth 1957, 181) as quoted in Evert, 2009) between words. It is the measure of how strongly one word predicts the other, given their individual frequencies and the observed frequency of their combination. (Evert, 2009) grouped PMI with effect size measures, which ask "how much does observed co-occurrence frequency exceed expected frequency?".

The standard *Gensim* implementation used the NPMI formula shown in 2.1

$$NPMI = \frac{ln(p(word_a, word_b)/(p(word_a) * p(word_b)))}{-ln(p(word_a, word_b)} \quad \text{where probability of word:} \quad p(word) = \frac{word_count}{corpus_word_count} \tag{2.1}$$

T-scorer was custom-made and computed the score based on the formula suggested by Gries (2010). Unlike (N)PMI, t-score belongs to the category of significance tests. This score indicated how unlikely was the null hypothesis that the words were independent.

T-score measure was calculated using the formula 2.2 after Gries (2010).

$$tscore = ln(\frac{bigram_count - expected}{\sqrt[2]{expected}} \quad \text{where:} \quad expected = \frac{worda_count * wordb_count}{corpus_word_count} \tag{2.2}$$

Both association measures are computed with minimum n-gram count set to 10. They use natural logarithm to the base $e$.

In both measures 0 means independence of n-gram components. NPMI lies in the [-1: 1] range, where -1 signals complete independence for words that never occur together, and +1 means perfect correlation. T-score bounds were experimentally established within [-11: 9] scope. To train the models, we set the association score threshold to the lower bound of each metric.

While learning *Phraser* models, we allowed for intervening words from the functional word classes, including adpositions, coordinating conjunctions, determiners, numerals and possessive pronouns. It gave access to items like *альфа::и::омега* (alpha::and::omega), *пакт::о::ненападение* (non::aggression::pact), *вкривъ::и::вкось*.

Then, the models were applied to the documents in parallel and unseen non-transition subcorpora.

### 2.3.4 N-gram Language Modelling Features

The second set of 10 lexical features exploits n-gram frequency lists and n-gram language models that were build from the same independent corpus resources as collocational features. These features reflect the distribution of n-grams in translated and non-translated texts, and replicate some of the QuEst++ blackbox features (Shah et al., 2013).

Word rank is a common translationese indicator which captures changes in the frequency profiles of lexical items in translated language. Translations are expected to use more of more frequent words, on the one hand, and more of *hapax legomena* and very low frequency items or out-of-vocabulary (OOV) words, i.e. 'strange strings' unseen in large non-translated corpora, on the other hand.

The implementations include measuring the size of the 'list head', covering top 1% of the frequency list from respective corpora, proposed by Laviosa (1998), mean word rank used by Volansky, Ordan and Wintner (2015), percentage of n-gram in each of frequency quartiles in a corpus of the source or target language (features 1046-1057) for sentence- and document-level quality estimation (QE) (Scarton, 2016). Sominsky and Wintner (2019) extended this approach to cross-lingual setting and compared the number of words in each of the seven bins defined in pre-trained frequency lists using word alignments.

Another popular approach to capture translational specificity at lexical level is using the measures of confusion from non-translations LM when it is presented with translated texts. LM probabilities and perplexities are included in the pool of 17 baseline features in QuEst++ (features 1009-1014). (Nikolaev et al., 2020) used predictability of translations based on LM entropy measures to explore the impact of the typological distance between SL and TT on the properties of translations.

This research relies on the following features:

- percentage of n-grams in the top frequency quartile defined on a large corpus of the respective language;
- percentage of n-grams in the bottom frequency quartile defined on a large corpus of the respective language;

- percentage of n-grams not seen in pre-trained frequency lists;
- LM perplexity averaged for document sentences;
- standard deviation for LM perplexities across sentences in the document.

N-gram frequency lists (of orders 1, 2 and 3) from the training corpus, including calculation of frequency quartiles, were generated using QuEst++ feature extraction module, following the instructions in Specia, Scarton and Specia (2016). These features capture the overuse of TL high-frequency items and, possibly, a higher ratio of OOV items. We used a trigram LM learnt on this corpus with *KenLM* library (Heafield, 2011) to generate average sentence perplexities and their standard deviation for each document in our data. We hypothesise that this model should return higher perplexities for translated language. N-gram features were generated from lemmatised and PoS-tagged version of the training and test corpora ('lempos' format), where words were represented with their lemmas disambiguated with PoS tags (e.g. Are work suits on the way out? -> be_AUX work_NOUN suit_NOUN on_ADP the_DET way_NOUN out_ADV ?_PUNCT).

## 2.4 Summary

Our work is related to the studies showing that translations stand apart in the TL and tend to share a set of lexical, syntactic and/ or textual features that can be cumulatively called translationese. We bring together findings from corpus-based translation studies and translation pedagogy, contrastive linguistics, register analysis and learner language research to suggest a set of translationese indicators that would effectively distinguish translations from non-translations and might be predictive of human translation quality and professionalism.

We defined translationese as a set of distinctive statistical characteristics of translations that set them apart from non-translations in the same register and make them an independent language variety. Although the professional norm in some registers might demand more covert translation than others, there is ample evidence that translations are always distinguishable from non-translations regardless of any external parameters, including of level of competence. Translationese is viewed as a by-product of the translation process, and it constitutes an unalienable property of all translations, the property

of being a translation.

A high-level theoretical explanation for the specificity of translations, translationese universals hypotheses, was faced with lack of empirical support, but yielded a consolidated understanding that translations bare traces of the source language as one of their strong predictors. There are also persistent characteristics that make translations their own type of language production distinct from non-translations, such as longer sentences, more connectives and clauses per sentence, lower TTR and other patterns that are neither TL nor SL oriented.

The link between the amount of translationese manifested in a document and its translation quality has been often assumed or implied in translationese studies, as well as directly explored in corpus-based setups. Although translationese is related only to the fluency aspect of quality, there are good reasons to believe that, in human assessment, quality judgments are highly influenced by the overall acceptability and readability of a translation, i.e. by how well it blends in with the expected TL norm. It seems intuitively reasonable to assume that the easier a text is recognised as a translation, the lower the quality.

The final part of the chapter presents three sets of translationese indicators selected and implemented taking into account the research goal and previous observations in a number of related research fields.

The outcomes of the experiments reported in this thesis largely depend on whether useful features were included in the document representation in the first place.

First and foremost, the features were selected to capture the specificity in the linguistic makeup of out-of-English Russian translations anticipated in practical translation textbooks and indicative of translation quality and level of professional expertise. Many of them reflect the typological differences between English and Russian. Feature selection was guided by other principles such as cross-linguistic comparability, reliability of extraction, frequency, interpretability and ability to abstract away from the content.

We proposed an extended set of morphosyntactic features extracted from UD annotation and two smaller sets of abstract lexical features that capture collocational and distributional properties of translations.

# Chapter 3 Human Translation Varieties

In essence, this chapter presents the textual data used in the experiments and previous work related to the study of professionalism in translation. This research relies on a hoist of resources, most importantly two self-compiled comparable parallel subcorpora containing student and professional translations of mass-media texts, which we refer to as *human translation varieties*.

Section 3.1 discusses previous work related to:

- building and exploiting LTC;
- defining professional levels and
- building comparable corpora of non-translations in translationese studies.

Section 3.2 gives details on the research corpus in this project and presents other language resources used to generate the features.

## 3.1 Competence Levels

Research into (human) translation quality requires annotation. Lack of manually annotated data is an acknowledged bottleneck for computational approaches in many research areas. Researchers are faced with a choice of either using existing resources, setting up an annotation experiment, or finding other natural sources of suitable labels for documents.

In the area of human translation quality estimation, professional and student translations, i.e. texts labelled according to the known professional status of translators, can be used as an approximation of quality labels. Note that some researchers prefer to speak of expertise levels rather than quality in this case.

Alternatively, researchers resort to human quality judgments passed in real-life situations where translations are routinely assessed or annotated. One obvious type of activity which generates human translation quality judgments is translator education.

An example of re-purposing the outputs of education process for research is presented in Wisniewski, Kübler and Yvon (2014). They produced a corpus of error-annotated machine translations from students' exercises done as part of their professional translator

training. Students on translation programmes are routinely asked to analyse translation errors and to suggest edits for identified faults. The authors describe a way to convert the outputs of this practice into a reliable machine-readable corpus, containing almost five thousand source sentences aligned with their machine translations and post-edited versions as well as error annotations according to a simple 6-category error classification and semi-structured qualitative analysis.

Another potential source of quality-annotated translations is the assessment procedures adopted within many translation degree programmes. Thus, collections of student translations are not only a source of documents with the known ontological status (usually compared to professional or MT translations as other members of the opposition). They often come with error annotation or metadata from other forms of assessment that can be used to generate fine-grained human translation quality labels and scores.

### 3.1.1 Learner Translator Corpora

The advent of available corpus building technology enabled the researchers to convert error-annotated and graded student translations into research corpora and gave rise to a number of corpus-building initiatives. This paragraph contains a description of typical practices in building learner translator corpora with a focus on error annotation, which is their main feature.

Following the inspiring example of *MeLLANGE* (Castagnoli et al., 2011)[1], there appeared half a dozen of learner translator corpora projects (Wurm, 2013; Lapshinova-Koltunski, 2013; Štěpánková, 2014; Espunya, 2014; Kutuzov and Kunilovskaya, 2014; Fictumova, 2017), many of which included error annotation. More recent additions to this family of corpora are described in Granger and Lefer (2018); Alfuraih (2019). The former presents *MUST* project and deserves a special mention for being a large-scale multilingual inter-university project, led by the Louvain Centre for English Corpus Linguistics with an extended expertise in corpus-based learner language research. In their annotation scheme, dubbed 'translation-oriented annotation', they lay special emphasis on the description of translator solutions: they have *Translation procedures* along with the traditional categories that reflect *ST-TT transfer* and *Language* categories.

Most of learner translator corpus projects rely on the outputs of translation-teaching

[1] corpus.leeds.ac.uk/mellange/about_mellange.html

frameworks as their source of error annotation. The choice of error taxonomy underlying a learner translator corpus or an annotation project covering student translations is governed by studies and observations from translator education, teaching procedures in place and criteria for TQA.

One of the early and comprehensive overviews of human translation assessment methods in educational contexts is offered by Secara (2005). That work also contained a description of *MeLLANGE* error typology, which was used in subsequent learner translator projects (Štěpánková, 2014; Fictumova, 2017; Verplaetse and Lambrechts, 2019).

Typically, error classifications designed for annotating student translations include two broad, theoretically-justified categories based on Toury's principles of adequacy and acceptability (Toury, 1995):

- content errors, i.e. "those which misrepresent the meaning (or style) of the source" (Chesterman, 2010, p.39), and reflect the relations between source and target texts (adequacy/accuracy aspect of quality). The subcategories of content errors sometimes include pragmatic sub-types, dealing with functional and communicative aspects of texts. For example, in *UPF Learner Translation Corpus*, this error category covers solutions that "compromise content, quality of linguistic and cultural expression and suitability for the purposes of the translation" (Espunya, 2014, p.36).
- language errors that reflect readability (fluency) and linguistic 'well-formedness' of the translated text; these errors can be revealed through comparison with non-translations in the TL.

Learner translator corpora can be multi-parallel, i.e. they include a number of translations of the same ST by many students. Besides, it is typical to maintain some register balance in these resources. For example, *VARTRA* includes student translations of texts in seven registers: political essays, fiction, manuals, popular-scientific articles, letters to share-holders, political essays and touristic leaflets (Lapshinova-Koltunski, 2013). It is not uncommon for LTC to store extensive metadata about students' gender, mother tongue (L1), level of the second language, level of education, situation of translation, etc (Bowker and Bennison, 2003, see, for example,). As it will be shown in Section 3.2.1 these parameters can be used to extract a more homogeneous sample of translations for experiments.

However, the exploitation of these translation-error-annotated corpora remains tricky: there are few publications that shows how the information retrieved from error annotation is applied in education process (the major intended use) or for any other purpose that goes beyond individual case studies and error tag statistics, typically released with the corpus.

In some projects, annotated translations are made available to students as a form of feedback, error statistics are used for tracking and analysing students' individual progress, and corpus data is used for generating exercises and tests (Kutuzov and Kunilovskaya, 2014; Fictumova, 2017).

Wurm (2020) used several hundreds of annotated translations to test two assumptions on translation competence acquisition. She compared error statistics for groups of translator trainees with different backgrounds and grades to see whether trainees' profiles correlated with the translation quality. She also measured effects of intensive training on some aspects of students' performance, including quality and speed of translation.

Kübler, Mestivier and Pecman (2018) described a teaching framework for specialised translation training, which was centred on using comparable specialised corpora in the translation process and included a learner component used to evaluate the outputs. The authors reported error analysis results for learner translations produced in different scenarios (with and without specialised corpora). They found clear evidence in favour of corpus use during translation production. The number of errors when using a corpus was lower for many error types. In particular, cases of *terms translated by a non-term, incorrect collocation, incorrect choice and preposition errors* went down. The paper suggested that error analysis could be informative with regard to SL units and language features that were particularly problematic for students. One example given by the authors linked many cases of "distortion" errors to the erroneous analysis of complex noun phrases (NPs). However, this direction of research was not developed: the authors focused on presenting corpus-based activities designed to address the most salient error types such as incorrect collocation.

There are several limitations for using error-tagged outputs of education process in a research project. Even if error annotation is implemented to produce the machine-readable output (which is not always the case), human errors are much more varied than those in MT and defy broad generalisations, especially given that the amount of annotated data

is sometimes limited to just a few texts of one type (i.e. *MeLLANGE* project annotated 232 student translations across four different text types and several language pairs).

Then, there are always issues with the reliability and validity of the annotations produced outside properly controlled research experiment settings. In educational contexts, it is not uncommon that annotation is performed by just one teacher.

Another limitation of a typical error annotation setup is annotation on the translation side of the parallel corpus only. This makes it difficult to identify the source language item or phenomenon that triggered the error. Arguably, identifying the factors that increase the error rate in translations, including the error-prone source text items, is one of the main purposes of error analysis.

One of the objectives in this thesis is to establish which assessment approach is best aligned with linguistic properties of translations by investigating the correlation between various quality labels and scores from a typical learner translator corpus and a given document representation. We contend that this is one way to validate human quality judgments generated in real-life contexts. Additionally, Section 4.4 reports the results of three inter-annotator reliability studies that demonstrate the level of agreement between humans in translation assessment task in various annotation setups.

### 3.1.2 Professionalism and Quality

Student and professional translations can be viewed as forming a continuous scale reflecting competence and expertise levels. These ontological categories offer objective labels for machine learning experiments, provided that the criteria of professionalism are clearly defined and other important parameters of the communicative situation with translation, such as register and mode, are controlled for as much as possible.

Ideally, students and professional translations should come from the same source texts like in *VARTRA* (Lapshinova-Koltunski, 2013) or in a study by Carl and Buch-Kromann (2010) who used translations from 12 professionals and 12 students for two short texts specifically obtained for the their study. However, in many cases researches rely on register-comparable collections of human translations obtained from different groups of subjects. Popovic (2020) used comparable ST for language pairs where the same ST for translators from different groups was unavailable. To minimise the effects of this setup, she suggested

*relative difference* between ST and TT as values for their features (sentence length, word length, morphosyntactic variety, lexical variety and lexical density). The results confirmed that there were learnable differences between translations produced at the different levels of expertise.

**Defining the categories** There are some acknowledged problems with delineating the two types of translators. Admittedly, students can vary in terms of how much translation practice they have been through. Tirkkonen-Condit (1990) compares the output of students in their first and fifth years of professional degree. The students can vary with respect of their mother tongue and mastery of the other language. Obviously, it is best to have a homogeneous group of students with regard to the socio-demographic factors relevant in translation. In a study by Daems et al. (2017), student translation were obtained from MA students of translation who had passed their final English Translation examination but had no professional experience beyond their studies. The group of professional translators in Daems et al. (2017) included people with a minimum of 5 years of full-time employment. The information about the years of experience can be unavailable when translations are obtained from real-world environments rather that within a specifically designed experimental setup involving specific compensated subjects[2]. In this case it is typical to assume that translations run by well-established reputable publishers were produced by experts doing their paid professional job (see Redelinghuys and Kruger, 2015, for example). Additionally, published professional translations usually carry translators' names and/or are endorsed by the editorial board. This is the approach taken in this thesis.

**Process- and product-oriented approaches to professional varieties** The studies that involve comparison of professional and student translations can be focused on either process-related or product-related aspects of translation. Translation quality focus can be viewed as a subtype of product-oriented studies.

Previous research focused on *procedural aspects* established differences between students and professionals in the overall time spent on translation and in the use of external

[2]Interestingly, Daems et al. (2017) acknowledged that professional translational behaviour can be affected by the experimental setup when translators do not perform routine tasks. This is another supporting argument in favour of using real-life translations.

resources such as dictionaries and reference materials as well as in the frequency and types of problem-solving activities. Professionals were usually faster and made less use of dictionaries. They had a tendency to translate at the higher linguistic levels than students and took into account coherence and register (Tirkkonen-Condit, 1990).

A study of professional varieties by Carl and Buch-Kromann (2010) combined process- and product-based approaches, with the latter being focused on quality. The authors were interested in the procedural differences between students and professionals. They employed eye-tracking and keystroke logging to look into the length of translation process phases (skimming, drafting and post-editing). They found that students spend more time familiarising themselves with the source text at the skimming stage and less time editing their translations than professionals. They also compared the quality of translations produced by students and professionals. For that they had all translations blindly evaluated by a native speaker of the TL for accuracy and fluency using a 5-point scale for scoring. The finding that is particularly relevant to this research is that "students and professionals differ mainly with respect to produced translation fluency" (Carl and Buch-Kromann, 2010).

In a similar vein, (Daems et al., 2017) employed eye-tracking and keystroke logging to explore translation speed, cognitive load, and the use of external resources for students and professional translators. The product comparison was quality-centred and used the results of manual error annotation. The authors observed that "students perform somewhat worse than professional translators, but the effect was not statistically significant" (p.258). At the same time, they found that the level of domain specialisation (general texts in their experiment) was negatively correlated with the number of errors. These results are not surprising, given that their experiment was staged to include students who had just completed their programme of professional education and general domain was the most common text type in their training. The study demonstrated that professionalism understood as an ability to produce successful translations and expertise understood as years of professional practice can be distinct from each other as translator's characteristics[3].

With regard to quality, early research results, referred to in Daems et al. (2017), were not conclusive whether students performed worse than professionals. However, if profes-

[3]We do not make this distinction in this dissertation.

sionalism in translation could be reliably linked to systemic language patterns (possibly, related to translationese), then professional translations could be used to work around the scarcity and unreliability of the data annotated for translation quality.

**Translationese perspective on professionalism** A fruitful type of product-oriented studies of professional varieties is built around the theory of translation universals and the concept of translational language. Some researchers within this strand make an explicit link between translationese-related properties and translation quality.

Corpas Pastor et al. (2008) and Ilisei (2012) used professional and student translations in medical domain to test the validity of simplification and convergence. They did not focus on the professional status of translators as a factor as such, but made marginal observations on the differences between the two translation varieties. For example, Corpas Pastor et al. (2008) noted that simplification hypothesis could not be confirmed for student translations, while some simplification features returned significant results for professional translations.

This approach is developed in two papers by Redelinghuys and Kruger (2015); Redelinghuys (2016). They aimed to establish whether the linguistic makeup of student and professional translations exhibited traces of the trends in expected translational behaviour (explicitation, normalisation, simplification and levelling-out of register differences). The authors hypothesised that experienced translators were more aware of the cognitive constraints of the translation process and had enough skill to counteract them. Professionals were expected to be more familiar with text conventions and translation norms. The results of the study included some evidence in favour of these universals, especially simplification, for all varieties of translations. However, the expected differences between subcorpora representing different levels of professionalism were often statistically insignificant, which was attributed to the small corpus size and its heterogeneous content in terms of SLs, varieties of TL (English) and registers. The overall conclusion substantiated our experimental setup: some of frequency-based features (conjunctive markers, *that* complementiser, standardised type-token ratio, word length, readability index) could potentially be considered indicators of translation expertise.

Only marginal differences between student and professional translations of the same English source texts were established in a number of publications by Lapshinova-Koltunski.

Those studies were based on a register-balanced parallel corpus, specifically designed to study professional varieties of translation (Lapshinova-Koltunski, 2013). The authors employed a number of computational techniques and representation approaches to explore the importance of expertise in shaping the properties of translations, including register. For example, Rubino, Lapshinova-Koltunski and Van Genabith (2016) resorted to supervised machine learning as a research method in an attempt to capture the differences between student and professional translations. They tested usefulness of features inspired by research in MTQE in a text categorisation task. The feature set included 13 surface features such as number of upper-cased letters, and over 700 surprisal and distortion features from regular and backward language models. However, their result for the binary professional/student translation classification was barely above 50% baseline demonstrating that features designed for MT were not helpful for that task. Lapshinova-Koltunski (2017) employed hierarchical cluster analysis to understand which factor in translation – register or translation method (categorised as student, professional and machine translation) – correlated with the selected frequency features more. They concluded that register explained variation in translation better. At the same time, in some registers the level of expertise proved to be more prominent than register. Bizzoni and Lapshinova-Koltunski (2021) used the same textual data to test information theory approach to document representation. For each translation they calculated perplexity from a neural LM trained on PoS sequences in non-translations. Contrary to expectations, professional translations generated higher perplexity scores, i.e. they were more difficult for the model to predict. They authors explained it by higher register variation in professional translations signalling higher sensitivity to register among professional translators.

In our own previous research, comparable student and professional translations of mass-media texts were teased apart on a morphosyntactic feature set (Kunilovskaya and Kutuzov, 2018). A binary classifier for the two varieties achieved F1-score of 72% on top 10 most useful features. That result was lower than for PoS trigrams as features, but helped to identify patterns that distinguished the two professional categories in English-to-Russian translation, including *nsubj:pass*, *xcomp* and *acl:relcl*, which were also the strongest translationese indicators.

In Kunilovskaya and Lapshinova-Koltunski (2020), the experiments were extended to include English-to-German language pair. That extension required adaptations to the

feature set to make the features comparable across the three languages. The results of translationese classifications (translations vs. non-translations) for each professional variety and statistical feature analyses indicated that the amount and types of translationese might be related to the level of expertise. For example, it was established that in both language pairs students produced more translationese related to shining-through. The difference between student and professional translations was machine-learnable: the binary classification on the full feature set returned 70% accuracy for the German translations, and 74% accuracy for Russian. This work also proposed an approach to capture the amount of translationese as Euclidean distances between the three text categories: aligned source and target texts, and comparable non-translations. In a visual representation, the shape of resulting triangles showed relative importance of shining-through, normalisation and SL/TL-independent translationese. That methodology was subsequently used to explore register as a factor in professional English-to-Russian translation (Kunilovskaya and Pastor, 2021).

**Professionalism as quality** A explicit link between professional varieties, translationese and quality was made by De Sutter et al. (2017). They represented quality as professional competence levels, comparing student and professional translations on two datasets: English-to-French translations of fiction and French-to-Dutch translation of news texts. In both case-studies student translations were defined as translations by first or second year students on profession-oriented MA programmes in two universities produced as part of their course work. All students had TL as their L1. Professional translations included published translated novels or translations from an existing parallel corpus. Importantly, they treated professional translations as a professional norm, demonstrating (with PCA plots and ANOVA analysis) that while professional translations did not entirely blend with comparable non-translations in the target language, students were usually further from the expected TL norm. The differences observed between professional and student translations were not clear-cut and "only seven features (out of 30) exhibit a significant difference between students and professionals" (p.33) in their first case study, for example. The authors contended that acceptability could be measured as distance to the TL convention represented in the linguistic behaviour of the professional translators and professional writers.

In summary, investigations into professionalism in translation is an established research area. The differences between two varieties are revealed on procedural level and in terms of linguistic properties of translations. The latter aspect encompasses direct comparisons of annotated quality between the categories, and comparisons of various translationese trends. The distance between a translation and the expected TL norm manifested in comparable non-translations is sometimes interpreted as an indicator of quality (with an eye on the observed professional standard). Generally, professional translators demonstrated similar though less obvious trends in their linguistic behaviour and were closer to non-translations. Automatic detection of professional varieties is notoriously difficult, especially when they come from the same source text. The results for differences in quality are at best inconclusive, not the least because most studies obtain student translations from advanced students, who are well-familiar with translation theory and have completed some practical translation modules. As shown in Popovic (2020), when human translation is produced by a true layman, e.g. a crowd-worker with unknown professional status, the difference between professional varieties is more detectable. Qualitative and quantitative description of professional translation subcorpus is given in Section 3.2.2.

### 3.1.3 Expected Target Language Norm

The second category in a traditional binary translationese classification is originally-authored text in the TL. Chesterman (2004) used the term *non-translations* to refer to that component of the research corpus. It is also known as a *reference corpus*, along with the other reference collection of source texts sometimes involved in translationese research. According to Chesterman (2004), non-translations are used the expected TL norm/standard to evaluate target text family fit (textual fit).

The concept of comparability of language resources is approached differently in (corpus-based) translation studies, on the one hand, and NLP, on the other.

NLP interprets comparable corpora as texts in the same topical domain. They are harvested on a set of seed terms (Kilgarriff, Avinesh and Pomikálek, 2011), and comparability is calculated based on the lexical features, such as vocabulary overlap or bag-of-words representations (Li, Gaussier and Yang, 2018). In many NLP tasks, functionally comparable bilingual corpora (such as the reference corpora of non-translations in translationese

studies) can be classed as quasi-comparable (see, for example, "non-aligned, and non-translated bilingual documents that could either be on the same topic (in-topic) or not (off-topic)" in Fung and Cheung, 2004, p.1051).

However, for researchers in corpus linguistics, translation studies and contrastive analysis, the same corpora are "comparable corpora in a strict sense" (McEnery and Xiao, 2007, p.19). In translation studies, functional, communicative and pragmatic adequacy is placed at the top of the hierarchy of translation quality criteria (Nord, 2006). In a general case, translations are expected to retain the source text functional properties and to conform to conventions of the respective text type in the TL. This makes the considerations of cross-linguistic functional comparability of the two reference corpora (source texts and non-translations in the TL) of paramount importance.

It is well-known that different **registers/genres trigger different type of translationese**. Lapshinova-Koltunski (2017) demonstrated that register is one of the major factors explaining variation in translation. Neumann (2013) revealed the specificity of German-English translations observed in some registers but not others. Delaere (2015) provided evidence that **translational norms were register-specific** and some registers could be more tolerant to interference and less prone to normalisation than others. The importance of building an adequate reference corpus is also reflected in the fact that some corpora (like *CroCo* and *Europarl-uds*) that are designed for translationese or contrastive research, include untranslated reference texts as their integral part (see Hansen-Schirra, Steiner and Neumann, 2012; Karakanta, Vela and Teich, 2018, respectively).

This research interprets comparability as a functional and register-related property, similarly to how bilingual comparable corpus is described in Kutuzov et al. (2016), or how it is traditionally defined in corpus-based translation studies (Zanettin, 1998). In particular, Zanettin (1998) lists subject domain, author-reader relationship, text origin and constitution, factuality, technicality and intended outcome (i.e. communicative function) among criteria for *comparable* corpus collection.

The functional and situational properties of texts are most immediately captured in the concepts of genre and register.

According to Lee (2001), *register* is a text-internal approach to text categorisation. It is based on the assumption that competent speakers align their linguistic choices with the communicative situation and use specific linguistic means to achieve their communicative

goals. Operationally, this approach relies on the statistical analysis of the lexicogrammatic features such as subordinating conjunctions, infinitives, discourse markers or demonstrative pronouns. The description of *genres* as text categories is built around extra-linguistic criteria, ranging from domain to the perceived speaker's goal and social roles of the parties involved. Genres are conventionally recognised communicative templates enabled under the appropriate sociocultural circumstances to achieve specific communicative purposes. They provide a more flexible and practical, although be it more subjective, approach to explain the real-life text variety.

These views are complementary: text categories delineated on extra-linguistic parameters are expected to share linguistic properties.

Besides, calculating frequencies of tokens (lexis-based categorisation typical for domain-oriented approach), can be as effective in genre classification as more elaborate register features. Xiao and McEnery (2005) showed that keywords analysis performed on par with Biber's features (Biber, 1988) in detecting both similar (everyday conversation vs official speech) and distant genres (spoken genres vs. academic prose). Their analysis demonstrated that the important differences in frequencies had to do with the function words, which formed the top of their keyness statistics results. It was shown that a set of 'register features' – as small as frequencies of nouns and frequencies of pronouns – could be effectively used to capture four major text categories in the BNC (Lijffijt and Nevalainen, 2017). Another important register feature is the distribution of connectives and other cohesive devices. Kunz et al. (2017, p.286) reported "considerable variation across registers, language internally. The differences between registers in German were more pronounced than in English". Note that features useful in register analysis were also among important translationese indicators discussed in Section 2.1.2. There have been numerous attempts to establish a link between genres and their linguistic features, while ignoring domain differences inside genre categories (including Lee (2001) and Braslavski (2010)). In this thesis, we prefer the term *register* which is more in line with the analysis of linguistic properties of texts, our key research interest.

The comparability of datasets and resources is usually ensured by collecting texts from similar sources (e.g. the same institutions, websites, or corpora), and by using the same chronological and sociolinguistic sampling frame. Alternatively, researchers can rely on the pre-existing register/genre annotation. However, even if the labels in two

given corpora coincide, it does not warrant the data comparability as has been repeatedly shown in the literature (Sharoff, 2018; Delaere, 2015). Sometimes, the description of resources comparability is limited to a phrase such as 'the BNC sample was chosen so as to mirror the makeup of the TEC' or 'reference corpus made comparable to the parallel data in terms of register'. The assumed comparability of monolingual and cross-lingual resources is typically a point of criticism. For example, in his overview of research on explicitation, Becher (2011) questioned the comparability of materials used in a number of cases. 'Applying the same sampling frame' (McEnery and Xiao, 2007) is contingent on the the analyst's recognition of the similarity between circumstances under which the texts were produced: Zanettin (1998) defines comparable corpora as put together on the basis of similarity of content, domain and communicative function as *perceived by a corpus compiler*.

On top of that, pre-defined categories pigeonhole texts in accordance with the accepted convention in a given language community, and does not allow for a more flexible and realistic reflection of the evolving text-type variety or for reliable cross-linguistic comparisons. Simple solutions, which work for the major text categories, fail in the presence of more subtle distinctions. For example, we found that the impressive and reproducible results from Lijffijt and Nevalainen (2017), where they achieved F1 = 90% in the classification of four 'tried and tested' top-level categories from British National Corpus (BNC) using pairs of simple register features like frequencies of nouns and pronouns, was reduced to only F1 = 71% on a less balanced six categories subcorpus described in Kunilovskaya and Sharoff (2019).

Building register-comparable corpora is treated as an important research step in this thesis. To this end, we proposed a solution based on representing texts with functional vectors and comparing texts on these representations. The vectors were produced by a recurrent neural network model trained on the hand-annotated data in English and Russian from the Functional Text Dimensions (FTD) project. More details are provided in Section 3.2.3.

## 3.2 Data Description

This section has a description and quantitative parameters of textual data used in this thesis. To achieve the goal of this research and to test whether translationese indicators can be used to predict translation quality, a corpus needs to be composed of three text types: source texts, target texts and comparable non-translations in the TL. Extraction of lexical features also requires independent language resources to train language models.

Training and evaluation of quality estimation models is performed on labels derived from the known professional status of translators (students vs. professionals) and on three types of manual quality scores assigned to student translations at sentence and document levels in a number of annotation projects. However, this section discusses the latter only in as much as they define the size of respective subcorpora within student translations. Detailed description of annotation experiments setup and results is provided in Section 4.4.

We view document as our main granularity level. Throughout this thesis, *document* refers to a set of contiguous sentences in their true order. The number of sentences in a document varies across the subcorpora, but it is not less than 10.

### 3.2.1 RusLTC Subsets

The primary component of our research corpus – various subsets of student translations – comes from the *Russian Learner Translator Corpus* (RusLTC) (Kutuzov and Kunilovskaya, 2014).

**RusLTC description** It is a large parallel bidirectional corpus of student translations in the English-Russian language pair. The project was conceptualised in 2012 at the *Translation Studies Department, University of Tyumen*, by translation teachers and translation degree students. The corpus collects outputs of professional translation degree programmes from 14 universities in Russia, however, most translations are by students with the University of Tyumen. It is an ongoing project, and new translations are being added after every semester. We have an effective preprocessing pipeline to integrate new students' submissions into the corpus. The most labour-intensive step in this process is manual correction of automatic sentence-level alignment. The alignment is done with *hunalign* library (Varga et al., 2007) and then manually corrected in *Heartsome TMX*

*Editor*, version 8[4]. Aligned bitexts are stored in Translation Memory eXchange (TMX) format, a subtype of Extensible Markup Language (XML). The source and translation segments are identified by corresponding attributes. In TMX, they retain the associated metadata, including links to original plain text files, which allows to retrieve the whole text if needed. Originally, TMX format is intended to store translations in several languages ('translation unit variants'). In *RusLTC*, this functionality is used to organise translations by several students.

Although the corpus is bidirectional, the English-to-Russian direction is much more represented. At the time of writing, RusLTC counts over 450 English source texts and over 4,000 of their Russian translations. The number of translations per source ranges from 1 to more than 60 and averages at 8. In the most typical scenario, students translated an extract of about 400 words. Students who donated their translations to the corpus have Russian as their mother tongue. At the time of translation, they would have been through at least three full years of English training at the university level, their English competence is assessed as B2 level.

Translations are produced as part of routine and exam assignments or as submissions to various translation contests. There are, however, translations from trainees who study translation as a supplementary course or study translation part-time. We also include translations from internship tasks by students majoring in translation and as graduation translation projects by part-time students. These student populations are described in the searchable corpus metadata. All translations are fully anonymised. Additionally, students' informed consent was obtained for including their translations in the corpus. The corpus website details a mechanism to withdraw their work from the corpus at any time.

The key extra-linguistic parameters related to the subjects and the situation of translation, discussed above, are reflected in 10 metadata fields:

1. Student's gender
2. Educational programme and mode
3. Year of study
4. Student's affiliation

[4] github.com/heartsome/tmxeditor8

5. Register
6. Completion (draft/final)[5]
7. Grade for translation
8. Independence (routine/exam/contest)
9. Conditions (home/classroom)
10. Year of production

Translations were made under different conditions related to the amount of stress and responsibility and time limit. Routine translations are less independent, because the source text would typically be discussed in the classroom before submission. In a general case scenario, no restrictions on use of reference materials or access to the Internet was imposed, and the students did not use any special computer-assisted technology tools. In terms of the translation task, the students were expected to return *a translation of publishable quality*, and the target audience of the translations was said to be more or less equal to the target audience of the source text. We did not include students' output for translation exercises that might be offered at sentence-level or be based on manipulated source texts to facilitate learning or to achieve a translation outcome beyond functional translation (e.g. adaptation, summative translation).

*RusLTC* includes translations of sources classified into the following eleven registers: academic, informational, essay, interview, tech, fiction, educational, speech, letters, advertisement, review. The distribution of the documents across the registers is heavily skewed towards informational and argumentative texts published by general interest mass media. Around 80% of the English source documents fall within this category. The English source texts for student translations were published in 2001-2016 by well-known English mass-media outlets, newspapers or magazines like *BBC*, *The Guardian*, *The USA Today*, *The New York Times*, *the Economist*, *Popular Mechanics*, *The Independent*.

When receiving the contributions from other universities, we checked whether translations comply with these requirements.

The grades for translation come from the assessment results implemented according to the procedures in each university. These are official judgments about translation quality announced to students and kept in the university records. The corpus converts various

[5]There was a plan to collect drafts and final versions to study revision in human translation, but we failed to systematically acquire this material

assessment scales to a simple 4-point scale, the most typical approach in Russian higher education, where 2 is the lowest mark and 5 is the highest mark. This information is available for over a third of all translations.

*RusLTC* has an online search interface[6]. The search engine installed on the project website allows building and saving multi-parallel concordances for tokens, lemmas or PoS tags. There are various options for filtering search results by context or by metadata parameters. However, the primary use of the corpus seems to be as textual data for computational studies of student translations.

The corpus is available for download as a single customised TMX file and as an archive of plain text files. The corpus in plain text format has *FileDocument* structure; each document has a header file with respective metadata. Sources, targets and metadata files are linked to each other by filename conventions. For example, *RU_1_2_6.txt* is Russian translation number 6 to an English source text in *EN_1_2.txt*. The respective metadata is stored as *RU_1_2_6.head.txt* and *EN_1_2.head.txt*.

Over 750 translations (300,000 words) were annotated for translation errors as part of feedback to students in a practical translation course over a period of six years (2013–2018). Error annotation was mostly done for independent translations carried out in the exam or competition contexts. Technically, the annotation was performed in the *brat* online environment (Stenetorp et al., 2012)[7], which allowed easy annotator collaboration and sharing the annotated texts with the students. Details on error annotation, including error typology, results of inter-rater reliability study and generation of error-based quality scores are reported in Section 4.4.2.

**Sampling** This thesis relies on various *RusLTC* subsets. Sampling is necessary to ensure homogeneity of textual data in terms of register, translation conditions, translators' professional status and type of quality assessment. Figure 3.1 puts the four extracted subsets (solid coloured circles) into perspective of all available materials, which helps to explain the rationale behind the selection. The size of each circle is proportional to the number of translated texts in each category. The numbers in brackets indicate the number of English source texts, translations and aligned sentences in ST-TT pairs.

[6] www.rus-ltc.org
[7] brat.nlplab.org

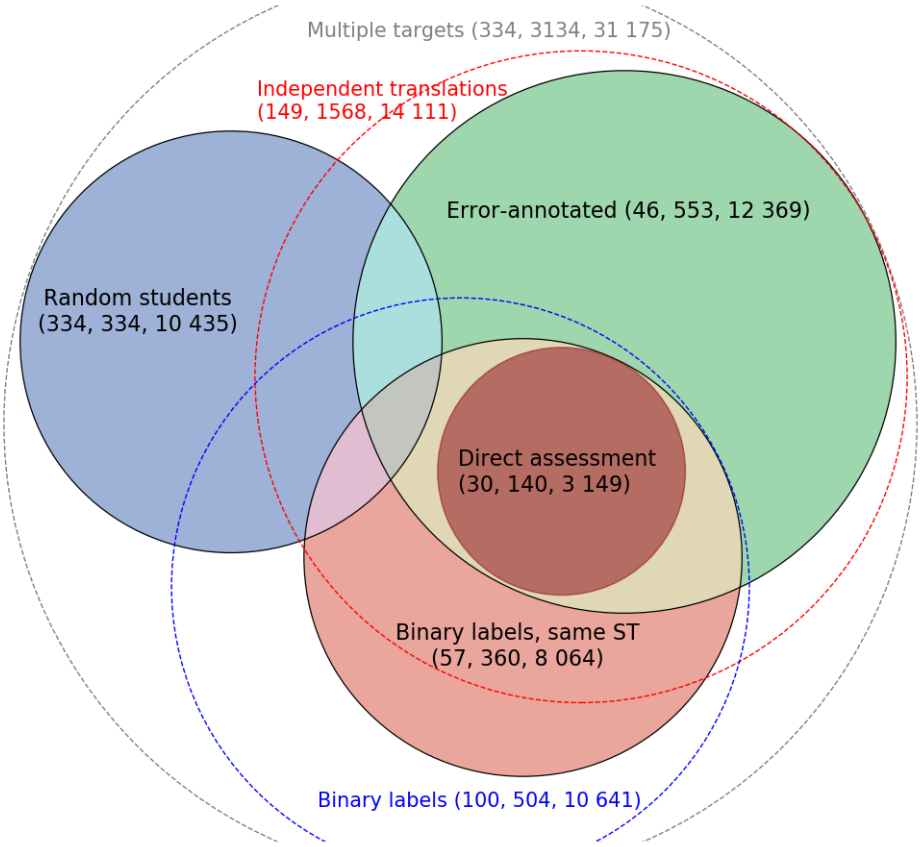



Figure 3.1: Number of (sources, targets, aligned sentences) by type of annotation

Note that for the purposes of this research we could not rely on the default *RusLTC* multi-parallel TMX, because our feature extraction setup required plain text documents aligned at sentence-level. We found that it was impossible to reliably restore respective source and multiple target texts from TMX as it did not retain the order of sentences in a document. Even though TMX is not designed to retain document structure, it usually does[8]. However, in our case, standard TMX underwent additional processing needed to put several translations of a segment into one translation unit as alternative translation unit variants, which interfered with the true order of sentences. Besides, there was lack of uniformity in sentence segmentation (sentence tokenisation) for each ST-TT pair amplified by lack of one-to-one correspondence between source and target sentences due to sentence splitting and merging operations. This resulted in loss of some sentences in our multi-parallel TMX as compared to plain text collection.

[8]Lack of context is identified as a current problem in MT training. Voita, Sennrich and Titov (2019) observed that in most practical scenarios, the sentence-aligned parallel data did not consist of complete documents undermining the development of context-aware NMT

To guarantee true sentence-level alignment for each ST-TT pair, we had to fall back to pairwise alignment of each document pair, leading to multiplication of source text files with slight differences in the number of sentences. As far as we can tell, this did not have any implications for our experimental setup.

Figure 3.1 shows that we had four divergent samples, each with its own type of quality label. They are comparable as there is some overlap between them, and they come from the same general population, e.g. *Multiple targets*, shown as a grey dashed circle encompassing other subsets. This general pool of student translations was selected according to the following parameters:

1. English-to-Russian translation direction;
2. produced by full-time students enrolled in specialist (5-year) or BA (4-year) degree programmes during their per-ultimate and final years of study;
3. ST is annotated for register as *informational, essay, educational*;
4. ST has more than 400 running words and at least 10 sentences.

These criteria, as well as *RusLTC* procedures contribute to the homogeneity of student component of the research corpus.

The *Random students* translation set (blue circle) is a subset that includes one random translation from all translations available for each of the 334 English sources selected for this project. While the multiple nature of *RusLTC* can be useful for some studies (see Castagnoli, 2009), it is certainly a factor to be considered when producing keyword lists and assessing frequency distributions of lexical items at a subcorpus level. This set was used to represent students ('stu') in comparison with professionals in experiments on professionalism in Section 6.1.1. The quantitive parameters are given in Table 3.1, along the parameters of the professional subcorpus.

The *Error-annotated* subset (green circle) includes independently produced translations that have complete error annotations. It means that translations were produced in an exam or competition setting, and the source text or draft translations were not discussed in the classroom prior to translation. This is a multiple translation parallel subcorpus for 46 English sources and 553 annotated translation, counting over 12 thousand sentence pairs in total. Admittedly, the number of unique English sentences in this subset is a bit over 1,000.

The *Binary labels* subset (light-brown circle), includes *good* and *bad* translations to the same source text. It is a ST-limited version of a larger collection (blue dashed line) that was build using the *grade* attribute. The binary quality labels (*good*/*bad*) are assigned to 1–3 top-ranking translations and to 1–3 bottom translations, respectively.

Finally, the *Direct assessment* subset (dark-brown circle) represents the texts annotated specifically for this thesis in 2020 in a direct assessment setup. More information on this is given in Section 4.4.3. Here it would suffice to say that the documents were selected to include other types of quality labels, as shown in Figure 3.1.

### 3.2.2 Professional Translations

Professional translations included in this research, were published by established electronic media, such as *BBC Russian service* (49%), InoSMI.ru (20%) and other electronic mass media (*Nezavisimaya Gazeta*, Russian editions of global mass media outlets: *Forbes* and *National Geographic*). These editions publish links to the respective source texts. A considerable portion of the professional subcorpus was harvested automatically by parsing the html code of the respective websites. All translations on these websites have the translator's name and/or are authorised by the editorial board. The automatically crawled data was filtered to exclude noisy results. Remaining 23% were sampled from the Russian National Corpus (RNC) parallel subcorpus of newspaper texts (Sitchinava, 2019).

To make sure that English sources for student and professional translations are register-comparable, we applied the methodology described in Kunilovskaya and Sharoff (2019). In brief, this approach consists (i) in vectorising documents with a ML model trained to predict document functionality along 10 dimensions and then (ii) in extracting the most similar documents from given text collections.

We relied on the most successful model from our previous research – a bidirectional LSTM with an attention layer trained on texts in mixed token-PoS representations suggested by Baroni and Bernardini (2006) (biLSTMmix). This experimental setup returned better results in the evaluation study predicting the functionality of documents in a corpus with 'known' register composition. This model was trained in a multi-label task. It used annotations described in detail in Sharoff (2018) and based on *Functional Text Dimensions* approach (ibid). In that annotation effort, linguistics students assigned texts a score

on a 4-point Likert scale that reflected the proximity of a text to the suggested functional prototype (argumentative, evaluative, informational, fiction, instruction, news, personal, promotional, technical-scientific, legal). The annotations were produced for English and Russian; the reported inter-rater agreement averaged at was reported at Krippendorff's $\alpha > 0.76$.

The quantitative parameters of the resulting comparable subcorpora of professional and student translations used in this thesis are reported in Table 3.1. For consistency, the table includes information about a comparable reference corpus of non-translations described in the next section (Section 3.2.3).

| | documents | sentences | tokens | |
|---|---|---|---|---|
| | | | EN | RU |
| students | 334 | 10,332 | 237,755 | 216,057 |
| professional | 404 | 17,008 | 421,514 | 383,983 |
| non-translations | 497 | 26,038 | – | 523,737 |

Table 3.1: Parallel subcorpora of professional and student translations and non-translated reference subcorpus

Note that we report the corpora sizes after filtering, pre-processing and annotation, based on the *CoNLL-U* format of the corpora, which was used to extract morphosyntactic translationese features.

Similar to student subcorpus, the automatic sentence alignment for professional translations was manually corrected using the same tools.

### 3.2.3 Non-translations Corpora and other Resources

This thesis uses two types of monolingual document collections comparable to the parallel data: (i) non-translation, or reference subcorpus, which is an integral part of the research corpus and is directly used in all experiments, and (ii) large monolingual resources used to train language models required for the extraction of collocational and n-gram features. These two collections are non-intersecting.

Non-translated Russian texts ('ref', i.e. reference corpus) were extracted from the main RNC (2016 version, counting in total over 50 thousand documents across a number of registers). For extracting a comparable subset, we relied on the corpus metadata and the following sampling frame:

1. longer than 350 words;
2. intended for non-specialist adult audience;
3. marked as article or commentary;
4. marked as neutral of style;
5. created after 2003

The resulting sample counted 8,210 documents.

Further on, this subcorpus of documents originally-authored in the target language (Russian) was subset to get a cross-linguistically register-comparable sample to the joint collection of English sources from student and professional parallel subcorpora. We used the same approach as for homogenising English sources for professional and student collections, but in a cross-linguistic setting. Its effectiveness was demonstrated in Kunilovskaya, Taslimipoor and Ilyushchenya (2019).

We used the most successful neural model to generate functional vectors for the total of 738 unique English sources to find two most cross-linguistically similar Russian non-translated documents for each of the English original texts.

After removing duplicates and 10% of the document-length and sentence-length outliers, this sample was reduced to 497 documents.

As we explained above, this subcorpus represents the expected target language norm for the selected register, i.e. the current TL 'textual fit' for professional and student translations.

*Language resources* used to learn *Phraser* and *n-gram* models to extract 24 collgram and 10 n-gram features were collected to represent mass-media register in Kunilovskaya and Pastor (2021). For Russian we used 790-million-word contemporary Russian newspaper corpus (2.7 million documents), included in the RNC monolingual resources. After filtering out short texts ($<$ 350 words), regional and specialised press using the standard corpus metadata, it was reduced to 226,144-document collection (see Table 3.2).

| | documents | sentences | tokens |
|---|---|---|---|
| en | - | 413,955 | 9,824,216 |
| ru | 226 K | 6,874,028 | 129,106,708 |

Table 3.2: Language resources for collocational, LM and *QuEst++* features

English models were trained on the BNC newspaper texts extracted following Lee's register annotation (Lee, 2001).

Note that the mass-media items in the BNC do not observe true document boundaries, but are in fact text chunks of varying length. However, it is irrelevant for the purposes of building LMs and n-gram lists.

### 3.2.4 Preprocessing

All text collections went through the same preprocessing pipeline. We deleted 10% of the document-length and sentence-length outliers, unified spelling to meet the conventions of the parser model (e.g. replaced various possible styles of quotes and double quotes with one variant: '«...»' to '"..."'), looked through frequency word lists to detect possible formatting issues, checked the number of sentences in parallel sentence-aligned documents, cleaned out remaining fragments of html code.

After sentence tokenisation (to get *LineSentence* corpus format), all lines shorter than 3 words were deleted (usually by-lines, dates and short headlines), as well as sentences starting with a lowercase character (obvious formatting errors or lists (see similar filtering rule in Guzmán et al., 2019).

Linguistic annotation was performed by *UDpipe* (Straka and Straková, 2017, v.1.2.0) pipeline (word tokenisation, POS-tagging, lemmatisation, parsing) with the language models pre-trained on English Web Treebank (EWT) (v.2.5) and *SynTagRus* (v2.5, Droganova, Lyashevskaya and Zeman, 2018) treebank for English and Russian, respectively. According to the UD description and performance reports[9] these are the largest treebanks available for the respective languages which outperform models trained on alternative annotated resources.

[9] ufal.mff.cuni.cz/udpipe/1/models#universal_dependencies_25_models

The documents in the resulting *CoNLL-U* format were used as input to the morphosyntactic feature extraction module.

## 3.3 Summary

This chapter describes all textual materials used in this thesis, their intended use and the rationale behind selection criteria. In the first part of the chapter we presented research related to the construction and use of learner translator corpora in (applied) research. This helped us to demonstrate that *RusLTC*, the source of student translations in this research, is representative of this type of language resources. It is a large parallel corpus with extensive metadata, including quality labels from real-life student translation assessment. We presented the sampling frame used to subset the RusLTC content to get reasonably homogeneous textual datasets for experiments and reduce the impact of possible confounding factors.

The chapter introduced the concept of human translation varieties: it describes approaches to defining student and professional translations and previous results related to their comparison with regard to (i) translators' practices and (ii) linguistic properties of the output, including quality. The latter line of research is either based on manual quality assessment (using Likert scale or error annotation) or on translationese features to see whether students generate more translationese and deviate from the expected TL norm to a greater extent than professional. The overall results indicated that differences in quality scores were small and mostly contained to the fluency aspect. However, student and professional translations have detectable differences in their linguistic makeup, with students being further away from the TL norm. De Sutter et al. (2017) interpreted these deviations as a quality indicator.

Finally, this chapter presents an approach to build register-comparable corpora for translationese research. We view this task as a necessary preliminary step, external to this thesis. Corpus-building efforts aimed to ensure (i) comparability of English sources in student and professional subcorpora and (ii) comparability of all English sources to Russian non-translations. To this end, the documents in both pairs were represented as vectors of 10 meaningful functional dimensions. Then, those vectors were used to extract most similar documents in the each pair of subcorpora.

Besides, all raw materials went through the same preprocessing step, including spelling standardisation and document-length and sentence-length normalisation, and linguistic annotation, using UD framework.

# Chapter 4 Quality in Translation

This chapter contextualises our approach to automatic generation of quality scores for human translations. It introduces key concepts related to translation quality and brings together the approaches to capturing quality in two disciplines that have a mutual interest in the subject: Translation Studies (TS) and machine translation.

The purpose of this part is two-fold: to overview automatic methods to predict quality of human and machine translation, and to describe the nature of the quality labels and scores obtained for this research, with references to recent best practices in benchmarking translation quality.

The chapter opens with a concise theoretical discussion of quality and its aspects, which provides the interpretations of key terms accepted in this project (Section 4.1). We explore the relation between human and machine translation, and implications for quality research regarding possible cross-disciplinary insights and techniques.

There are two prominent approaches to produce automatic quality scores in MT: *evaluation* and *estimation*. The more general term *assessment* is associated with manual appraisal of translations and creating annotated datasets required to test evaluation metrics and estimation methods. We contend that *quality estimation* is a more suitable approach for automatic prediction of human translation quality. Therefore, Section 4.2 has a focus on numeric representations and learning methods used to solve Translation Quality Estimation (TQE) task for MT. Importantly, we overview existing research dealing with automatic human translation quality estimation, a line of studies this thesis seeks to contribute to.

Section 4.3 presents most popular experimental setups to collect human quality judgments and produce reliable quality benchmarks. It discusses available options, standards and requirements with regard to the key parameters of quality annotation: experimental unit, raters, granularity of assessment, quality control measures and reliability studies. This description is used as a springboard to report setups and results of three quality annotation procedures which generated various quality judgments for student translations underpinning this dissertation (see Section 4.4).

## 4.1 Quality and its Aspects

Over the past few decades, the controversies about translation quality in MT and TS seem to have converged to a few general statements that are universally accepted at least theoretically. In both fields there is a wide consensus that the definition of quality is relative in several dimensions. First and foremost, it is contingent on the purpose of evaluation. It also varies depending on the communicative situation of translation and with regard to other available translation candidates.

The goals of translation evaluation can include diagnostics and system optimisation, comparative analysis or producing quality benchmarks. In each scenario the task can be set with regard to some pre-defined aspects of quality or quality standards. In this research we are mostly concerned with measuring the *overall end-user quality of translation as a product*. We treat human and machine translations as translation varieties that can in principle be evaluated within the same paradigm. However, the practical usefulness of a universal approach is doubtful, especially if the evaluation goals go beyond summative scoring, but aim at revealing the weaker components of the translation process to facilitate improvements (in either human or machine performance).

This overview abstains from purely theoretical expositions, following the principle that "the proof of the pudding is in the eating". In empirical studies the concept of quality can be defined through operationalisations, i.e. it depends on the nature of features/representations, and on how the labels were generated. However, we cannot avoid a brief discussion of translation quality aspects (Section 4.1.1).

Section 4.1.2 explores the commonalities and differences observed in approaches to quality in human and machine translation, which reflect the differences in how quality is conceptualised in the two fields. This helps to define the intersection between machine and human translation modelling techniques and their cross-disciplinary applicability. Section 4.3.1 discusses the intimate link between quality assessment for human translation and automatic quality predictions for MT. The quality of HT turned out to be an important factor in training and evaluating MT as discussed in a number of recent publications (see, for example, Popovic, 2020; Läubli et al., 2020). This aspect of HT/MT relation was brought into the fore after the introduction of NMT and a palpable improvement of automatic translation. Most importantly the quality of HT (including direction of

translation) was shown to have "practical impact on evaluation of machine translation (MT) systems" (Popovic, 2020, p.365). Popovic (2020) demonstrated that standard MT evaluation metrics returned different results when reference translations were produced by people with different professional status: professional translators, crowd contributors, or translation students.

### 4.1.1 Theoretical Underpinnings of Quality

This paragraph briefly outlines the translation-theoretical concept of quality that, according to the literature, is shared by researchers working on human and on machine translation. We also highlight the key differences in the methods used to determine quality that are indicative of double standards in assessing translations produced by professional translators and automatically.

It is not uncommon for MT researchers to refer to TS theorists to validate their approach to measuring quality in MT. There is a shared and wide-spread understanding in both fields that quality is a relative category. We will focus on two parameters of relativity: it can be defined (i) as appropriateness of a translated text to fulfil a communicative purpose in a given communicative context and (ii) as a relation to other possible renditions of a source text under given translation conditions.

The first dimension of relativity is conceptualised around the category of *adequacy*, which is the ultimate quality requirement: all translation solutions should, first of all, be acceptable from the functional-communicative-pragmatic point of view. From the quality research point of view, it means that the question of *'Is it a good translation?* is replaced with the question *'Is this an acceptable translation, given its communicative purpose?'* *Adequacy* is complemented with *accuracy* and *fluency* to get the triad of key aspects of quality, where accuracy can be briefly defined as the degree of meaning preservation, and fluency as the adherence to TL norms. For example, see the recognition of this triad and the hierarchy with regard to HT in Chesterman (1998) and to MT quality in Koponen (2010).

There is a known inherent controversy in the interrelations between adequacy and the other aspects of quality. While adequacy is a super-ordinate predominant category and can justify some deviations from accuracy and fluency, there are certain limits to the scope of source text modifications that can be done in the name of adequacy. These limits

are set by the need to preserve the source text identity in translation, which is a fuzzy and complex idea, forming the core of translation equivalence theories (such as proposed by Nida, 1964). If the key parameters of the prospective TT communicative situation are very different from those of the ST (e.g. different text functionality or audience), the scale of required adaptations might ruin the ability of TT to represent the ST, which is incompatible with the idea of translation (naively put as 'the same in the other language').

With regard to the audiences, some researchers attempt more practical explanations of how much alteration is allowed for an instance of a cross-linguistic transfer to fall into the category of translation, and not cross-linguistic adaptation. Latyshev and Semenov (2003) restricts the transformations possible in translation to differences in the linguistic and cultural components of communicative competences of the source and target audiences. This motivates replacements for culture-specific phenomena or explanations added for them in translation. For example, in *I packed my two Gladstones*, a typically British name for a leather suitcase, can be replaced by a generic description that omits *Gladstones* to avoid the readers' confusion over an insignificant detail.

In the most typical translation task communicative conditions (including the type of audience and text function) of the source and target texts are assumed comparable. Some researchers and teachers explicitly indicate this in translation briefs: "the target audience of the translations was said to be more or less equal to the target audience of the source text" (Daems, Macken and Vandepitte, 2013). Student translations used as research material in this dissertation were produced to solve a typical translation task: they were offered a mass-media text to be translated for information purposes to publishable quality standards.

The second dimension of relativity has to do with the space of possible renditions of a source segment acceptable under given translation task and conditions. In HT, it is customary to compare translation solutions with the available or hypothesised alternatives. This approach is called for by many translation educators. For example, Bittner (2020, p.172) states: "Good translation quality can only be better translation quality, just as bad translation quality can only be worse translation quality. There is no use dismissing a translation solution as unacceptable unless a better alternative can be produced."). Pym (2003) includes the ability to generate alternative translation solutions and to chose the best of them for a given context into the core of translator professional competence.

In an empirical setup, this means that a 'good' translation is not an absolute grade of quality, but the 'best' translation of all available. However, at a translation competition, a jury can agree that no translation is up to the expected standard and the winner is not selected.

In MT context, the notion of relative goodness, is most directly implemented in ranking, a form of direct assessment, where the participants are asked to place translations in the order of preference according to perceived quality.

### 4.1.2 Humans vs Machines

Despite TQA is an area of mutual interest for TS, a discipline in charge of HT, and MT, a discipline that develops automatic translation, there have been little interdisciplinary research between the two (Ahrenberg, 2017).

This lack of communication is manifested in some confusion over the use of terms, most notably, adequacy. MT-related descriptions use the term *adequacy* as synonymous to *accuracy* (see, for example, the definitions for Multidimensional Quality Metrics (MQM) framework[1]. Within functional translation theories (represented, for example, by Nord, 1997; Vermeer, 1989, who developed the *Scopos* theory by Reiss and Vermeer (1984)), a dominant paradigm in TS today, it is typical to make a distinction between *adequacy* and *accuracy*, with the priority of the former. Above all, a translation is expected to be fit for purpose of translation (translation task), i.e. be adequate to the extralinguistic context of translation, including intended audience, medium, domain, cultural conventions, etc.

By design this research is product-based and, hence, is set within the textual, more linguistically oriented, direction of TS (as opposed to cognitive, cultural and sociological branches, following disciplinary map of TS in Chesterman, 2005). This means that we rely on the concept of register, a theoretical construct focused on textual properties reflecting situation of communication, to capture adequacy in translation.

Judging by how quality is operationalised in the realms of human and machine translation, there is a gap between what is expected in terms of quality from a machine and from a human. This gap was revealed through critical analysis of the first claims in Hassan et al. (2018) that MT has achieved human parity. It turned out that the outcome of the comparison depends on how you measure quality. Läubli, Sennrich and Volk (2018) used

[1] www.qt21.eu/mqm-definition/definition-2015-12-30.html (Accessed: 31 July 2022)

alternative evaluation protocols to challenge the MT-human parity claim. In the years to come we might see this gap closing due to the urgent need for, and development of, new quality benchmarks and methods to adequately reflect the improvements of neural MT models. These new approaches address very 'human' aspects of quality (such as document-level coherence, for example). Nonetheless, at the time of writing, sentence-level reference-based quality metrics remain unchallenged in the MT research community.

**Key differences and their implications for quality research** This paragraph lists several factors that make some MT approaches to quality inapplicable to human translation.

First and foremost, TS discusses quality as a document-level phenomenon and HT are usually produced and assessed at document level, while MT models as well as most MT quality metrics *operate on the sentence level*. This is seen as one of the key shortcomings of the modern MT, including its neural implementations. There is an ongoing active research that aims to make neural MT more context-aware. The researchers have to design their own quality metrics to demonstrate achievements of the proposed models because the standard MT quality metrics (most notably, Bilingual Evaluation Understudy (BLEU)) are unable to capture improvements in MT on the text level. For example, Voita, Sennrich and Titov (2019) created a test set for targeted evaluation of improvements in deixis, ellipsis and lexical cohesion for their MT model that was additionally trained on document-level parallel data. They demonstrated that models indistinguishable with BLEU can be very different in terms of rendering text cohesion.

Läubli, Sennrich and Volk (2018, p.4791), who tested the claim of Chinese-English MT attaining human parity, emphasised "the need to shift towards document-level evaluation as machine translation improves to the degree that errors which are hard or impossible to spot at the sentence-level".

Document-level quality approaches in MT are still in the budding. The respective quality estimation shared task was introduced at WMT in 2016. The document scores to be predicted were obtained from a two-stage post-editing approach[2], which penalised documents that needed more editing when the context of the entire document was considered.

[2]It is inspired by the practice of outsourcing translation of sensitive documents as randomised sentences, and then double-checking re-assembled translated documents.

In 2018 a HTER[3]-based scores were replaced with quality scores from word-level error annotation. However, in both cases the document scores were generated from sentence-level edits or word-level errors with severity weights.

Interestingly, *QuEst++* document-level quality estimation method, the WMT baseline in 2016 and 2018, essentially used the same features as *QuEst++* sentence-level method. The values for those features (such as token length, ratios of four types of n-grams in top and bottom frequency quartiles, LM probabilities for each of the parallel texts) were averaged across sentences for document pairs. Scarton (2016) suggested nine features dubbed 'discourse-aware' in her PhD thesis. They were based on counting word repetition in ST and TT, and the ratios between those values[4]. In 2020 the baseline system computed MQM document-level scores using the word-level baseline system[5].

Next, the neural-based supervised machine learning approaches to quality estimation are held back by the lack of effective document representation relevant to translation quality. The current practice to produce document-level predictions by averaging sentence-level outputs seems sub-optimal.

Sentence-level is the most common setup for error annotation in MT, unless the researchers specifically target document-level phenomena like in Voita, Sennrich and Titov (2019); Läubli, Sennrich and Volk (2018). The sentence-level annotation setups for MT can be exemplified by *BLAST*, an early translation error annotation tool designed for "human error analysis of MT output, where humans identify and classify the problems in machine translated sentences" (Stymne, 2011, p.56) and by *ACCOLÉ*, a more recent MT-oriented platform designed to facilitate the task of "labelling translation errors by locating manually their occurrences in the target as well as in the source sentence" (Esperança-Rodier, Brunet-Manquat and Eady, 2019, p.28). Sentence-level setup is chosen in many MT studies that involve error-annotation (see, for example, Van Brussel, Tezcan and Macken, 2018).

The second most obvious distinction between human and machine translation which makes them substantially different text types (according to Specia, Scarton and Paetzold

[3]Human Translation Error Rate (HTER)

[4]e.g. DocLevelFeature1084: percentage of content words in the source document; DocLevelFeature9994: noun repetition in the target document, the number of words that repeat are normalised by the total number of content words

[5]statmt.org/wmt20/quality-estimation-task.html

(2018)) is the *degree of variation* observed. MT output is much more standardised in terms of number and properties of possible translation solutions. Ahrenberg (2017) in their analysis based on translation procedures (or shifts) showed that MT output was much more limited in the repertoire of microtextual transformations than HT of the same source text. This makes MT quality metrics based on a reference (known as the *translation quality evaluation* paradigm) not applicable to human translation. It has been repeatedly shown that the reference-based approaches, in particular, BLEU score, are unfairly constraint by design, even if they involve several references. Vela, Schumann and Wurm (2014) demonstrated that the reference-based MT metrics could not cope with the legitimate lexical variation typical for multiple human translations. The reference-based quality metrics favour compliance with standardised solutions, they penalise creativity and variety, the parameters that are often promoted as a key to success in translator education. Hence, we consider the *translation quality estimation* paradigm in MT quality research as theoretically better suited for human translation.

The greater variation observed in human translation explains *more detailed and flexible schemes and scales of quality* in TQA. TS has developed elaborate systems of aspects, criteria and categories of quality, most of which were excessive for MT before very recently. The granularity of analysis in terms of number of discernible aspects is a point of some debate is MT. On the one hand, there is evidence that accuracy-fluency distinction is purely analytical: human annotators of MT produced highly correlated scores for accuracy and fluency that led the researchers believe that the distinction might not be too practical (Callison-Burch et al., 2007). There is a trend to dispense with it in annotation setups acknowledged in a recent survey of MT evaluation metrics (Chatzikoumi, 2020). However, for translation teachers annotating students translations (after minimal training and calibrating sessions) the inter-rater agreement for the top-level error categories is above average, even though it is not uncommon to see the same error tag (such as text cohesion) included in either one or the other top-level category (content or language).

Third, in MT it is typical to distinguish translations that are expected to comply with *different quality standards*, ranging from *gist translation* that allows the user to identify texts of interest, and *publishable (production) quality* translations that maintain a delicate equilibrium between adequacy, accuracy and fluency to "satisfy the linguistic norms of a target culture and are adapted to the assumed knowledge of its readers" (Ahrenberg, 2017,

p.21). This distinction is made in Scarton (2016) and in Specia, Scarton and Paetzold (2018, p.43), for example. They present various scores as interpretations of quality suited for specific assessment purposes. Gisting and publication purposes are distinguished as laying various emphasis on fluency. Translation for publication is viewed as meant "to be read by someone who cannot speak the source language", with adequacy being critical. Interestingly, measuring quality as post-editing effort is outside of this line of reasoning. In this case quality is interpreted as ease of bringing MT to the required quality standard: some errors are easier to fix than others. Different quality expectations are reflected in the distinction between full and light post-editing of MT output. In the first case, the result is expected to be human-like, while in the second case it is characterised as 'good enough' (Massardo et al., 2016). In the context of TS, post-editing can be compared to the revision stage of translation. However, it is often presented as an integral part of a translation job, rather than an external operation.

On a cline between 'good enough for gisting' and publication quality, the goals for humans have traditionally been set higher (not only in terms of the end-product, but also in terms of the translation techniques used). MT is expected to be more literal, which is interpreted as semantic faithfulness. This property is counted upon in the pipelines where a MT system is fine-tuned to process a specific text type written in *controlled language* (with a limited range of the lexical and grammatical SL structures) (Gil and Pym, 2006, p.17).

Because source sentences can differ in the extent to which a literal translation is acceptable in TL, the outcome of human judgment about MT quality is contingent on the selection of sentences to be translated. This factor in MT assessment was acknowledged by the organisers of WMT-2020, who deliberately included challenging sources (short sentences with unconventional grammar, rich in colloquialisms) into the Russian-to-English shared task[6].

Another limitation to extend MT techniques to HT consists in the fact that state-of-the-art unsupervised approaches to MTQE rely on the internal information from the MT system (Fomicheva, Sun, Yankovskaya, Blain, Guzmán, Fishel, Aletras, Chaudhary and Specia, 2020) (known as *glass-box* or *confidence* features, also used in *QuEst++* feature sets), and are hardly usable until the parameters of human translation process can be

[6] statmt.org/wmt20/quality-estimation-task.html

reliably converted to numeric representation.

Finally, it might be useful to lay different emphasis on accuracy and fluency in HTQE and MTQE. It seems that major problems of unprofessional human translation (represented by student translations) and today's state-of-the-art neural translation are associated with different aspects of quality. (Carl and Buch-Kromann, 2010) demonstrated that translation students were "able to reproduce the source text meaning in their native target language just as well as professionals", but their renditions were less fluent. These finding were corroborated in our own investigation of SL-related issues in student translations (Kunilovskaya et al., 2022). This means that the distinctions between relatively high-quality human translations (professional translators and translation students) have to do with fluency. Modern neural MT is notorious for accuracy errors, especially omissions (see, for example, Van Brussel, Tezcan and Macken, 2018) and hallucinations, even though we might not mean the same degree of flexibility when we talk about machine and human accuracy.

## 4.2 Quality Estimation for MT and HT

### 4.2.1 QE as NLP Task

Quality estimation is one of two approaches to produce automatic quality scores for MT. Unlike a more traditional and well-established method to quantify quality in MT, known as *quality evaluation*[7], QE aims to score MT outputs in a reference-free manner, that is without comparing a candidate translation to a human translation or several human translations of a given source segment.

The major advantage of this approach over reference-based evaluation metrics is that QE can be used in online MT applications to score entirely unseen input, for which human reference translations are unavailable. QE is also designed to return results for individual data points, rather than to evaluate the overall MT system performance. It is better at capturing quality variation between outputs of the same MT system.

QE is often framed as a supervised ML task (classification or regression depending on the type of labels), which requires (i) explicit features or representations extracted

[7]Description of quality evaluation approaches is beyond the scope of this thesis

from texts and (ii) labels. According to Specia, Scarton and Paetzold (2018, p.1), "the general framework that is used to address QE of human texts can be similar to that for QE of automatic applications". They admit, however, that MT and HT are "substantially different" types of texts. Given the differences and the focus of this thesis, existing research in the area of MTQE and HTQE is summarised in separate subsections.

Note that unlike models that use specific pre-defined features as input to a learning algorithm, most recent achievements are obtained using syncretic feature-learning approaches. These models interpret text representation as a trainable (rather than frozen) layer of a modelling architecture. Neural feature-learning architectures are discussed in a separate paragraph below.

**QE levels** QE can be implemented at various levels: word, phrase, sentence and document. This section is focused on sentence and document levels as relevant for this thesis.

*Sentence-level* is by far the most popular setting. It is reasonable given that sentence is the main operational unit in MT systems. As shown above, in TS it is very uncommon to assess separate sentences, because human translation is essentially document-based. If any MTQE methods can be carried over to HTQE, they are likely to be document-based. Besides, human translations datasets are usually annotated at document-level. Even error annotation, which usually locates problematic spans within a sentence, reflects document-level quality because (i) a typical taxonomy for HT includes pragmatics- and texture-related categories, and (ii) annotators are exposed to the entire documents rather than isolated sentences and can annotate errors like tautology and repetitive content, for example.

In contrast to that, *document-level* MTQE research suffers from lack of annotated data. The first specialised shared task at WMT-2015 relied on automatic reference-based Metric for Evaluation of Translation with Explicit ORdering (METEOR) scores as labels. In 2016 they were replaced by scores from human two-stage post-editing with weighting to emphasise specific document-level problems. The dataset included 146 English-to-Spanish documents (news domain). The document-level QE shared task was re-introduced in 2018 and ran till 2020 on English-to-French NMT of Amazon product reviews. That dataset was error annotated by professional translators using MQM error typology for

accuracy, fluency and style errors (training set: 1,000 documents, 129,099 words). It is not clear whether the annotators had access to the whole document. The document-level scores were generated from error statistics weighted by error severities and normalised by document word count (Specia et al., 2018). This task attracted little attention from the community, and the results did not improve much over the years (based on the best-performing system): $r = 0.534$ (2018), $r = 0.374$ (2019)[8], $r = 0.573$ (2020). This approach might run counter the previous concept of document-level quality as being more than the simple aggregation of its sentence-level or word-level quality scores.

Specia, Scarton and Paetzold (2018) acknowledged that document-level QE is the least explored of the granularity levels. This task is faced with two major problems: (i) the majority of MT approaches are trained on parallel data with shuffled sentences and no possibility to recover documents, and (ii) obtaining human labels/scores, which are not aggregated from lower level annotations, is a challenge, not the least because there is a lack of parallel corpora with preserved document boundaries.

In the last two years document-level WMT QE task was replaced with critical error detection, a subtype of *word-level* QE based on error annotation rather than post-edits, and aimed at predicting parts of translation that are dangerously unreliable to alert post-editors. In this novel task a sentence receives a binary label depending on whether it contains a catastrophic error.

**Evaluation setup** QE evaluation setup depends on the type of labels used. QE tasks formulated as *classification problems on discrete labels* are evaluated on standard metrics such as accuracy, precision and recall.

For completeness of this description, let us define *accuracy* as the number of all correct predictions divided by the total number of instances in the dataset, *recall* as the number of correct positive predictions divided by the total number of instances belonging to the class (positives), and precision as the number of correct positive predictions divided by the total number of positive predictions. It is typical to report a harmonic mean of precision and recall, calculated as shown in Equation 4.1 and known as *F1 score*.

$$F1 = 2 * (precision * recall)/(precision + recall) \tag{4.1}$$

[8]according to oarganisers, the results are not comparable because 2019 used a different test set

In dataset construction and in experimental setup, we ensured the balance between number of instances in our classes, therefore we use macro-averaged F1 score, which calculates F1 score independently for each class and then takes the average treating all classes equally.

Our experimental Chapters 5 and 6 report accuracy, precision, recall and F1-scores as classifiers' performance measures, expressed as percentages, and standard deviation (STD) (a measure of variance) across 10 folds in a cross-validation setting.

Specia, Scarton and Paetzold (2018) recommended to supplement them with graphical representation of a classifier's performance with receiver operating characteristic (ROC) curves, which visualise the relation between sensitivity (the true positive rate) and specificity (the true negative rate) for different thresholds to find the optimal equilibrium between the highest true positive rate and the lowest true negative rate, represented by a curve reaching for the upper-left corner. To report this balance as a number, area under (ROC) curve (AUC) statistic is used. AUC values vary in the range from 0 to 1, where 1 is AUC for a perfect classifier.

For *continuous scores (learnt with regression models)*, it is common to use *Pearson's r correlation coefficient* proposed by (Graham, 2015) as a more reliable QE evaluation approach. This coefficient measures linear correlation between two variables and is estimated as the ratio of covariance of two variables to the product of their STD (see Equation 4.2).

The coefficient value lies in the range from -1 to 1, where a value of 1 represents a perfect positive relationship, -1 a perfect negative correlation, and 0 indicates the absence of a relationship between variables. In QE setting the variables are gold scores generated from human annotations and scores predicted by models for all test instances.

$$r = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2(y_i - \overline{y})^2}} \tag{4.2}$$

To control for variance-related effects, non-parametric Mean Absolute Error (MAE) and Root Mean Square Error (RMSE) are used. They do not make the assumption of normal distribution of values, and would penalise predictions for instances which deviate too much from gold standard.

A variant of Spearman's $\rho$, a rank-based version of Pearson's correlation coefficient, is used for WMT ranking tasks. It is a non-parametric test that aims to measure monotonicity of the relationship between two variables, given the difference in ranks between predicted and true labels.

Pairwise comparisons of participant systems predictions to human ranking was also measured as Kendall's $\tau$, i.e. the difference between concordant and discordant pairs normalised to total pairs.

### 4.2.2 Representations in MTQE

Like many other NLP tasks, MTQE has witnessed a shift from feature-based to neural approaches. From 2017-2018 most application-oriented research is based on feature-learning or fine-tuning neural architectures, including *POSTECH*, a predictor-estimator architecture (Kim et al., 2017), *DeepQuest* for document-level tasks (Ive, Blain and Specia, 2018), followed by a rapid succession of other neural frameworks for QE, including NeuralQE (Martins et al., 2017) and *TransQuest* (Ranasinghe, Orasan and Mitkov, 2020), the winner of the *WMT2020* competition in QE in sentence-level DA tasks.

While neural representations are more effective as means to an end and approximate observed PE effort or other human targets better than hand-engineered features, they lack a diagnostic potential. This motivates researcher to design alternative datasets and types of annotation to target specific problems in MT generation (see research by Scarton, 2016; Voita, Sennrich and Titov, 2019, mentioned above). WMT-2022 introduced a new error-annotation-based QE shared task – Explainable QE – where participants are requested "to infer translation errors as an explanation for sentence-level quality scores"[9] and return continuous token-level QE scores, with peaks corresponding to erroneous translation solutions. Generally, in the area of MT, most successful hand-engineered features, included in 17-feature *QuEst++* baseline, tend to reflect the surface of language rather than theoretic linguistic reasoning, which might be better aligned with the ontological status of MT.

In our opinion, broadly understood high-level explainability, linked to translation procedures and competence, is viewed as more important in TS than in MT. However,

[9] wmt-qe-task.github.io/subtasks/task2/

there were attempts to exploit linguistic language-specific features for MTQE task. Most notably, Kozlova, Shmatova and Frolov (2016) extracted features from parse trees of source and target sentences (number of dependencies from root node, proportion of internal nodes, number of subjects, relative clauses, etc). They studied the contribution of those features and their effectiveness compared to standard *QuEst++* set to demonstrate that this line of work has research potential.

Another limitation of feature-learning neural architectures is lack of required amount of labelled data. The WMT dataset counts seven thousand sentence pairs for training with further two thousands for development and testing. This might not be enough to learn representation from scratch, and most researchers rely on pre-trained embeddings in a fine-tuning, transfer learning or zero-shot learning setups.

**Handcrafted frequency features** The MTQE approaches rely on linguistic properties of sources and corresponding MT outputs (MT system-independent *black-box* features) as well as on internal information from respective MT systems (*glass-box* or *confidence* features). Note that the latter are not used at document level.

In *QuEst++* framework, a well-known QE pipeline, *black-box* features are further classified into three groups depending on whether they were extracted from sources, targets or both (Specia, Scarton and Paetzold, 2018).

In terms of granularity levels, document-level approaches mostly follow the work on sentence-level. The rest of this paragraph gives examples of the features that are proposed for document-level in *QuEst++*. We focus on 17 features included in *QuEst++* baseline at WMT before neural baselines were introduced. The full document-level feature set lists 78 features[10]. We used 64 of them in our experiments, excluding features with low variance and those dependent on sophisticated external resources. To be precise, we removed 14 features that required parsing (9300–9307), pre-defined lists for sentence intralingual triggers (1302–1304) or had low variance (1086, 1087, 1094).

*Complexity* features characterise difficulty of sources on the assumption that more complex sentences yield lower translation quality. They are extracted from source documents. The first seven items in the list below (comprising 13 features) are included in the baseline

[10] github.com/ghpaetzold/questplusplus/blob/master/config/features/features_doc-level_all.xml

and are shown in the order of *QuEst++* configuration file). The three features beyond that are taken from the extended feature set.

1. number of tokens;
2. LM log probability of n-grams in the source document using a large corpus of SL to build LM;
3. average source token length;
4. average number of translations per source word in the source document using probabilistic dictionaries (two features with different probability thresholds and weighting strategies);
5. percentage of unigrams/bigrams/trigrams in the document that fall in the first/fourth quartile of frequency in a large corpus of the source language (six features);
6. percentage of unigrams in the source document seen in a corpus (MT training corpus);
7. number of punctuation marks;
8. percentage of content words;
9. percentage of nouns;
10. average word frequency: on average, each type (unigram) in the source document appears x times in the corpus (in all quartiles);
11. etc.

Features from targets are known as *fluency* features. They measure "how natural, fluent, and grammatical the translation is" (Specia, Scarton and Paetzold, 2018, p.53). They are mostly the same as ST complexity features (except #4), but are evaluated on the target document. A few features that are not measured for the source document are listed below:

1. number of occurrences of the target word within the target hypothesis (averaged for all words in the hypothesis – type/token ratio)
2. perplexity of the target sentence (with and without end of sentence marker);
3. proportion of OOV tokens in the target document;
4. etc.

These specific features aim to capture the likelihood of translation to be seen in the TL.

Features from sentence-aligned sources and targets are called *adequacy* features. They assess the formal correspondence and semantic equivalence between source and target segments. Specia, Scarton and Paetzold (2018) highlights issues with extraction because these features are cross-linguistic by design. Most of them are expressed as simple ratios or differences between complexity and fluency features for aligned sentences, averaged to get document-level values. For example:

1. difference between number of periods in source and target (absolute and normalised by target length);
2. ratio of content word repetition between the target and source documents;
3. ratio of percentage of verbs in the source and target;
4. etc.

Discourse-aware features are limited to nine features that count repetition of tokens, lemmas and nouns in the document and their ratios.

These features remained a standard and were used as WMT QE shared tasks baseline for several years. They are also used as baseline in related research (e.g. Kozlova, Shmatova and Frolov, 2016, including for HTQE in Yuan, Sharoff and Babych (2016) and subsequent publications).

Various attempts to create proxies for reference translations or to use back-translations as features are outside the scope of this description as less relevant for HTQE.

**Distributed representations** The most simple approach to create numeric representations for documents from embedding models, both static and dynamic, is to average word vectors in the document, generating a single vector for either source or target document (see Specia, Scarton and Paetzold, 2018, p.73).

The descriptions in this paragraph are limited to the best-performing neural-architecture solutions for NMT as per the results of WMT for the past few years on DA scores for sentence-level and on MQM scores for document-level.

However, we start with *predictor-estimator* model that dominated the leader-board at WMT-2017 and can help to trace the evolution of QE methods based on distributed representations.

Predictor-estimator is an internally complex sequential model consisting of two components: a predictor, which generates embeddings for words based on the context in which they appear; and an estimator, which produces quality estimates (Kim, Lee and Na, 2017, POSTECH).

The first component is a multi-layer model. It produced three representations: one for the source sentence using a bidirectional recurrent neural networks (RNN) model with the attention layer and two embeddings for preceding and succeeding context of each word in the translation segment by forward and backward RNN layers. The outputs of these three layers are summed and further processed by *Maxout* layer which outputs pre- and post-prediction vectors. A concatenation of these vectors is sent to the estimator, a simple bi-directional RNN layer, which produces quality estimates. This model is used in a multi-task setup: it outputs word-, phrase-, and sentence-level quality labels.

The estimator and the predictor are connected using a technique called stacked propagation, which allows for the information learned by the estimator to be propagated back to the predictor.

Additionally, the WMT-2017 solution is derived from an averaged output from 15 model variants.

A dedicated document-level-architecture was proposed by Ive, Blain and Specia (2018) and is known as *DeepQuest*. It learns representations at sentence level and uses a bi-RNN encoder with an attention mechanism to learn weights for different sentences. Finally, a sigmoid output layer receives the weighted sum of sentence vectors and produces real-valued predictions for a document.

Starting from 2019, WMT uses *OpenKiwi* (Kepler et al., 2019) to implement various neural baselines, including *predictor-estimator* for sentence-level Direct Assessment (DA) task in 2020, for example. 2019 was the first year when all datasets were generated by NMT and QE models were evaluated against human judgment (DA, news domain), i.e. viewed as a metric.

Sentence-level HTER-score and document-level MQM-score baselines were produced by a neural quality estimation system (NuQE) (Martins et al., 2017). It represents the same class of solutions as *predictor-estimator*, except it relies on pre-trained embeddings as input, *NuQE* leverages word-level HTER/MQM annotations (BAD, OK) to obtain composite predictions on higher levels using an encoder-decoder architecture with bidirectional

RNN layers. According to organisers, many participants preferred to use contextual pre-trained embedding models such as BERT and XLM along with transfer learning and fine-tuning (Fonseca et al., 2019).

For example, one of the best performing approaches for sentence- and document-level tasks described in Yankovskaya, Tättar and Fishel (2019) did not learn representations from scratch like earlier approaches (e.g. *DeepQuest*), but used Bidirectional Encoder Representations from Transformers (BERT) and cross-lingual *LASER* embeddings as input to their neural architecture to predict either sentence HTER or document-level DA. They experimented with MT internal information as input (log probability) and with pre-training on synthetic output (an automatic MT metric score for all available WMT submissions).

Baseline solutions offered by the organisers for document-level DA task (QE as a metric) avoided complex learning architectures and produced predictions in an unsupervised manner. They were based on either (i) similarity scores between ST and TT vectorised with a cross-lingual model, or on (ii) overall log-probability of the translated document extracted from the respective NMT system (Fonseca et al., 2019). Other unsupervised approaches on MT-internal 'glass-box' information are reported in Fomicheva, Sun, Yankovskaya, Blain, Guzmán, Fishel, Aletras, Chaudhary and Specia (2020).

A similar approach is used by the WMT-2020 winner in sentence-level DA- and HTER-score tasks. *TransQuest* (Ranasinghe, Orasan and Mitkov, 2020) presented an ensemble of the two architectures, which rely on vectors from pre-trained XLM-RoBERTa (XLM-R) (Conneau et al., 2020)[11] embedding model as input: (i) the first architecture computes a pooled representation for CLS token from each pair of sentences and passes it as input to a softmax layer that predicts the quality score of the translation; (ii) the second architecture encodes the source sentence and its translation with two separate XLM-R transformer models. The values of cosine similarity between the two are output as the model predictions.

According to the organisers, in both WMT-2020 and 2021 (Specia et al., 2020, 2021) the key to success was the size of pre-trained multilingual embeddings, ensembling and data augmentation: the baseline models relied on the same architectures in essence, but lacked these features. Particularly high results for some language pairs are explained by

[11] huggingface.co/xlm-roberta-large

very low quality 'hallucinated' translations in these datasets, so, basically, by variability of MT output.

In short, feature-learning approaches are better than approaches on explicit features, but they are outperformed by glass-box methods, given the same amount of data. Pre-trained multilingual embeddings perform even better. Document-level QE remained at the periphery of researchers' interest; the respective task was removed from the WMT competition in 2021.

### 4.2.3 Human Translation Quality Estimation

Unlike in the area of MT, where the research into automatic prediction of translation quality is thriving, there are only a few attempts to extend this approach to human translation. Some of the reservations for such a cross-over are discussed in Section 4.1.2. Research in HTQE is also hampered by lack of enough annotated data to apply neural architectures. As shown above, feature-learning models are data-greedy. Even if a model relied on pre-trained embeddings in a transfer learning or fine-tuning scenario, around several thousand annotated instances are needed to achieve reasonable results. This is certainly a problem for any document-level setup, which motivates researchers to move to sentence level or to use unsupervised approaches.

To the best of our knowledge, there are two projects, that are directly relevant for the present thesis. Most notably, there is a series of experiments for English-to-Chinese human translation estimation reported by Yuan, Sharoff and Babych (2016); Yuan (2018); Yuan and Sharoff (2020). Yuan set out to compare feature-engineering and feature-learning approaches for QE task formulated as a regression problem on document and sentence level.

That work was based on 457 multiple student translations, counting 3,569 sentences in total to six English source texts, each from a distinct register from science to literature, ranging in length from 5 to 15 sentences (229 to 313 words) (Yuan, 2018). Translations were produced by third-year or fourth-year English major students and come from a parallel English-as-Foreign-Language learner corpus.

Student translations were annotated by two raters following a percentile scoring scheme according to the ATA *Certification Programme Rubric for Grading*, which had been used

in the ATA translation qualification certification test before 2015, when it was replaced with an error annotation scheme. The Rubric had four assessed dimensions:

1. **usefulness** the extent to which the meaning of the ST is conveyed;
2. **terminology** lexical appropriateness in terms of terminology, register and style;
3. **idiomatic writing** smoothness of the translation in the TL;
4. **target mechanics** adherence to the TL norms.

The first two dimensions were collapsed into an additional adequacy score (in the range of 1–60), and the last two – into a fluency score (1–40) to get a total quality score on a 1–100 scale.

The level of agreement between two raters in the experiment is reported at staggering Krippendorff's $\alpha$ between 0.74–0.96 across four dimensions for document- and sentence-level separately.

Yuan's (2018) document-level experiments are based on an extensive set of 360 language-independent surface features, including subsets of monolingual, bilingual and language modelling features, some of which are reminiscent of *QuEst++* approach and some are novel.

To give an impression of the types of information they leveraged, we will give a few examples of their explicit document-level features:

**monolingual** :

- dependency tags,
- semantic role labels,
- constituency parsing probabilities,
- cohesion features (such as normalised frequencies of various pronouns and connectives, pronoun/noun ratios),
- coherence features (lexical tightness: PMI of content words in a sliding window of -5:5, averaged cosine distances of adjacent sentences computed from vectors of bag-of-words);

**bilingual** :

- Manhattan distance or log ratios of monolingual features,

- MT back-translation similarity,
- similarity between ST and TT word vectors from cross-lingual embeddings,
- pseudo-reference agreement scores,
- word and phrase alignment features (normalised proportion of aligned words by ST length (word alignment precision), normalised proportion of aligned words by TT length (word alignment recall), summation of the logarithmic probability scores of aligned words),
- bilingual terminology.

**language modelling features** : LM probabilities/perplexities and OOV words

These features are fed to *XGBoost* algorithm with grid search parameter optimisation to get Pearson's $r = 0.72$ against *QuEst++* baseline result of $r = 0.67$ in document-level experiments.

For sentence-level HTQE, he used a stacked neural model using convolutional neural networks (CNN) and RNN layers with a cross-lingual attention mechanism. The proposed feature-learning model, initialised with pre-trained word embeddings, demonstrated second-best performance to the reduced set of most correlative explicit features proposed for sentence-level (see Table 7.3 Sentence-level HTQE results Yuan, 2018, p.160). The values for Pearson's coefficient are reported at $r = 0.53$ and $r = 0.42$ for feature-based and feature-learning approaches at sentence level, respectively.

We might want to treat these results with a grain of salt, however. Given that Yuan's translational data is very repetitive (more than 50 translations to one source), it made sense to adjust train-test splitting procedure (p.123) to respect groups in the data is all settings, so that training is done on translations to five sources and testing on the translations from the sixth set. Otherwise, the models might have learnt what is good and what is bad for translations of this particular ST, and the test set was not a problem for them.

We aim to test an alternative text representation approach on a more varied set of student translations in another language pair and with other types of labels/scores, but we roughly follow the same design: document-level experiments on frequency-based features, sentence-level experiments on embeddings.

Yuan demonstrated that hand-engineered features, apart from returning competitive

performance, had the advantage of interpretability. The best-performing features selected in their experiments helped determine grammatical characteristics that are likely to be responsible for lower translation quality scores. They showed that human translations typically contained errors beyond the lexical level, to which proximity-based MT evaluation metrics were less sensitive.

The second study, directly related to HTQE, is reported in Zhou and Bollegala (2019). They interpret translation quality as an unsupervised STS task with the view of detecting high/low quality human translations. In their approach, words in ST and TT are represented with cross-lingual word embeddings, obtained by aligning Japanese and English embedding spaces from *fastText* models in *MUSE*[12]. The authors experimental efforts are focused on identifying the best method to measure the distance between words in ST and TT. They came up with *Bidirectional Minimum Word Mover Distance* method, which matched most similar source and target words in a bidirectional way and returned the sum of two minimal cumulative costs for (i) translating each word in ST to all words in TT and (ii) in the opposite direction.

Their dataset consisted of 130 sentences randomly extracted from a larger set of sentences from Japanese manuals to digital cameras translated into English by 50 crowd-source native Japanese speakers who have studied English as a foreign language. Recruited translators had various levels of the English language and translation competence, which ensured a broad spectrum of quality levels in the dataset.

Each sentence pair in the dataset was graded by four bilinguals with extensive experience in practical translation using a scale from 0.0 to 1.0, with four quarterly descriptors for easy range-finding. The scores from four raters were averaged; no inter-rater agreement statistic was provided. The best proposed metric achieved Pearson's $r = 0.529$.

As was already mentioned, Rubino, Lapshinova-Koltunski and Van Genabith (2016) reported a study which tried to capture professionalism in English-to-German translation using a typical QE setup: surface features, log-probabilities trained on delexicalised PoS and flattened syntactic trees, backward LM features. Their work was motivated by findings in Aharoni, Koppel and Goldberg (2014) demonstrating that the higher the confidence of a translationese classifier that an instance is a translated text, the lower the translation quality. However, their data did not contain any quality labels, apart

[12] github.com/facebookresearch/MUSE

from translator's professional status, and their features did not perform well on the task, returning a classification result only marginally above a baseline.

## 4.3 TQA and Designing Quality Benchmarks

As seen from the previous discussion, TQA, a field which investigates and suggests methods to produce reliable and consistent human judgments about translations, can be viewed as theoretical premises for benchmarking quality in translation. This section summarises TQA practices applied to human translations in academic and industrial contexts, and overviews annotation setups used in MT to build translation quality benchmarks for testing metrics and QE approaches. Both fields are relevant for this dissertation because we use computational methods to investigate human translation quality.

### 4.3.1 Quality Assessment and Benchmarking in HT and MT

**HT: analytical vs holistic approaches** Quality assessment of human translation is usually set within one of two major paradigms: error-analytical or rubrics-holistic, with possible hybrid approaches. The former is based on annotation of specific spans in the text following an error taxonomy, and the latter is descriptor-based, where a translation is placed at a certain level of achievement according to each criterion included in the assessment scheme.

As we will show below, the modern implementations of both methods have scoring and descriptive functionality, and return very similar results. Rubrics enjoy comparable popularity in the educational sector and with translator accreditation bodies, while industry seems to prefer error-analytical approaches. Generally, it is believed that errors are more objective and product-oriented; they are preferable when some justification of the outcome is needed. Holistic approaches provide a better account of macrotextual and pragmatic issues of translation and offer an integral characteristic of a student's performance/competences against an expected standard; it might be better suited for formative assessment. Another difference is that error annotation is points-off system which is focused on negative aspects of translation, while rubrics give a positive perspective of achievement levels along specified dimensions.

Waddington (2001) compared the results returned within each approach for the same translations in a university exam setting. His statistical findings indicated in favour of error analysis in terms of internal consistency between raters, especially when applied to more homogeneous subgroups of bad and good translations. However, the author finds that the reliability estimates increase if two approaches are combined in a proportion of 70/30. Results reported by Eyckmans, Anckaert and Segers (2009) corroborate the conclusion about higher reliability of the analytic approach. Interestingly, the authors propose an alternative method, based on the calculation of discriminative power of items identified by raters as errors. A more recent research found that rubrics could be as reliable as error analysis (Turner, Lai and Huang, 2010).

One of the early and comprehensive overviews of human translation assessment methods offered in Secara (2005), demonstrated a general converging trend in educational and industrial approaches towards hierarchical error-based models with a manageable number of error types (about a few dozens) weighted on a pre-defined scale. Naturally, in translator education there is more emphasis on the descriptive potential of the systems to provide formative feedback to students, while professional translation agencies are more interested in a reliable summative metric. There is an established opinion that it is preferable to apply real-life (industrial) quality assessment strategies in vocational translator training (Doyle, 2003; Koby and Baer, 2005; Williams, 2013) even though product-centred and training-centred TQA might have different objects of assessment.

According to Hegrenæs and Simonnæs (2020), there is a continuing trend for educational and certifying institutions to move towards descriptor-based assessment methods, sometimes incorporating an analytical component. In the ATA certification examination final grade is obtained from error statistics, but the assessors are provided with a flowchart for grading to ensure that final result truthfully reflects the candidate's competence in four areas: *Target Language Mechanics, Meaning Transfer, Writing Ability*[13]. It looks like the choice between the two approaches for testing professional translators is largely determined by tradition (for example, the United Kingdom and Norwegian translating/interpreting accreditation examination systems use descriptors, while the ATA uses a well-known error taxonomy).

[13] www.atanet.org/certification/how-the-exam-is-graded/error-categories/ (Accessed: 14 July 2022)

Despite error-based approaches appeared as a result of growing demand for more objective and explicit criteria for translation description and TQA, most rigorously implemented schemes, such as the ATA assessment framework, still raise doubts as to their objectivity as demonstrated in a manual analysis of ten ATA assessors' outcomes and feedback by Phelan (2017).

In NLP setting a standard way to validate annotation results is to carry out a proper inter-rater reliability study. In TS settings, it is not uncommon to avoid rigorous statistic evaluation while comparing various TQA approaches. One of the exceptions to this tradition is the project by Colina (2008), who reported inter-rater reliability for the proposed functionalist/componential holistic assessment tool. Their reliability study results were quite encouraging: the authors recorded surprisingly high and consistent *Fisher-Z* values of over 0.8 (which converts to Pearson $r$ 0.66) obtained for three texts which were assessed by up to 19 assessors grouped by profile (bilinguals, professional translators, language teachers) across three language pairs. The raters were asked to select one of several statements that characterised compliance of translations with four criteria (*Target Language, Functional and Textual Adequacy, Non-Specialized Content Transfer, Specialized Content and Terminology*). For each of three texts each rater returned 4 scores. Unfortunately, the authors chose to focus on other aspects of the tool usefulness, such as users' survey, and did not comment on the selection of the statistical measure and how it was applied, except in a short footnote.

Apart from reported comparable reliability and popularity in the sector, errors and rubrics are often contrasted in their ability to capture the pragmatic aspect of translation quality. Its importance became widely acknowledged with the development of functionalist and communicative theories of translation in the late 1980s. Arguably, holistic approaches are better designed to account for text-level pragmatic factors (compliance with the translation brief, explicit or implicit client specifications). These factors are notorious for escaping explicit attention in error analysis. At the same time, theoretically convincing descriptive-explanatory models of translation quality (Nord, 2006; House, 2001) are criticized by Williams (2001) for lack of a measuring scale. This criticism was answered by the development of a number of rubric-based approaches. According to Hegrenæs and Simonnæs (2020), they were used as scoring tools to qualitatively assess translations, including in terms of extra-linguistic pragmatic factors. For example, Jiménez-Crespo (2009) tries to

reconcile the error-based approach typical in localisation industry and the communicative-pragmatic requirements to quality by proposing an additional holistic level to assess the entire text through descriptive statements. The research is motivated by the lack of account for the communicative-pragmatic factors in the standard error typology, while the adequacy objective is formulated in the *LISA* documentation as the requirement for localised texts and products to "look like [they have] been developed in-country" (LISA, 2004, p.11). Generally, error-based quality metrics, sometimes complemented with rubric descriptors, dominate among the assessment methods used by the translation providers for *translation assurance*.

Similarly, the state-of-the-art analytical framework used in MT assessment, the harmonised MQM allows overall translation assessment based on questions or statements, corresponding to the provided error (issue) types as well as on user-defined specifications, such as reader impression, sentiment, clarity, whether the translation enables a task to be completed (DQF-MQM, 2015).

One of the reasons to prefer analytical approaches over the holistic ones is to collect word- or phrase-level quality annotations for computerised analysis. Annotations that localise target text quality issues to particular sentences and their parts can be used to identify translationally most challenging source sentences and source language related translational problems to be focused in further training of both humans and machines (Popović, 2017; Kunilovskaya et al., 2022).

In summary, error-based and criteria-based quality assessment frameworks for HT were shown to deliver comparable results and reliability. They enjoy similar popularity in academia, certification and industry, and can be used to provide summative and descriptive results. Both approaches incorporate means to reflect pragmatic aspect of translation quality. HT assessment is typically performed on the document-level by translation experts, who tend to agree more in their judgments than participants in MT annotation experiments. It is not only the level of expertise and responsibility that explains this difference. We would argue that translation quality is better perceived at document level, when local context and global textual properties help to gain true understanding of how successful a translation is.

**Labels and scores for MT** The ultimate yardstick in evaluating any automatic system is human performance on the same task.

Interestingly, there is much more pessimism about getting human annotators agree in their quality judgment (especially at document-level) in the field of MT than in TS. Generalising from previous research, Specia, Scarton and Paetzold (2018, p.64) concludes that "document-level evaluation (in general) is less trivial than the evaluation of more fine-grained levels".

Lack of inter-rater agreement does not make TQA for MT unique among other tasks required from humans. It is not unusual for people to disagree on topic boundaries or types of the discourse units, or semantic similarity. Nonetheless, in the area of human judgments about MT quality, especially in the context of error annotation, low inter-annotator agreement is admitted to be endemic because "humans differ in their understanding of quality problems and their causes" due to the lack of single, objectively correct translation for any given text (Lommel, Popović and Burchardt, 2014). The challenge of producing reliable and accurate quality scores makes researchers devise more elaborate and well-defined ways of capturing human opinions, and fairer measures to demonstrate consistency and truthfulness of the annotations. This problem got a lot of traction quite recently when improvements in Neural Machine Translation (NMT) revealed common flaws in MT benchmarking practices.

Besides, in MT quality annotation is viewed as more *task-dependent*, which explains active exploration of various labels. Choosing the right yardstick remains a matter of continuous debate in MT. Which type of annotation is more feasible and reliable to produce scores as a machine learning target? Various annotation setups can to produce diverging scores because they essentially measure different properties of translation. (Yankovskaya, Tättar and Fishel, 2019, p.101) gives a convincing example demonstrating how DA scores can be different from HTER.

The uncertainty about an appropriate annotation procedure to produce scores for testing QE models is evident if we look at the changes in preferred annotation methodology for WMT datasets in the past few years.

In 2018, when NMT was first introduced into the QE competition, participants of the traditional sentence-level tasks were asked to score (and rank) sentences according to *post-editing effort*, predicting HTER, post-editing time in seconds, and counts of various

types of keystrokes. There was an obvious gap in the performance of QE methods on Statistical Machine Translation (SMT) and NMT output in the same language pair (e.g. the best results for English-German went from $r = 0.739$ to $r = 0.513$).

In the same edition of the competition document-level *error-based* scores on English-French NMT were introduced.

In 2020 and 2021 the organisers offered DA scores for their traditional sentence-level predictions task, while keeping HTER for word-level and sentence-level tasks on just two language pairs. MQM error annotations retained its positions as label of choice for document-level.

The upcoming 2022 competition[14] announced that error annotation is going to be used for the majority of tasks. To get an impression on how prediction results vary for sentence-level HTER and DA scores, consider English-German NMT 0.513 (2018, post-editing (PE)), 0.572 (2019, PE), 0.554 (2020, DA) and 0.584 (2021, DA). Note that there was also a switch of texts type from IT domain to Wikipedia texts. For English-Russian language pair, immediately relevant for this research, there are only two results available: 0.5923 (2019, PE on NMT, tech domain) and the staggering 0.808 (2020, DA) for Russian-English NMT on a challenging text type full of unconventional grammar and colloquialisms, which was likely to generate a favourable-for-ML variety in DA scores.

This paragraph presents a brief overview of manual scores used to learn and test models in MT quality research. To keep the description relevant to the purposes of this work, we omit the description of labels from ranking candidates generated by various MT systems and from ranking MT outputs in the same test set that were used as target in QE research in 2007-2016 (QE as a ranking task) and other approaches based on automatically generated labels, or labels from reading comprehension tests. Further on, approaches that suggest synthetic scores or manipulated input (artificially degraded quality) as a proxy for human judgment also fall outside of the scope of this work (see, for example, Saadany and Orasan, 2021, where they introduce modifications into MT output to capture critical errors related to sentiment transfer). For the sake of brevity, we will also leave out the discussion of post-editing as it is less relevant for human translation.

[14] wmt-qe-task.github.io/

**Error Annotation: converging practices** Error annotation is an established approach to produce quality scores for both machine and human translations. It is also particularly suited to get separate scores for the two major aspects of quality (accuracy/adequacy and fluency) in addition to the overall scores. Commonly, these scores are used to evaluate automatic quality metrics by measuring correlation between the algorithm predictions and the human judgments represented by the scores. For example, Ustaszewski (2019) used major error categories to get specific subsets of machine translated sentences, which contained or did not contain accuracy/adequacy or fluency errors, to test cross-lingual semantic similarity between the source sentences and their targets of different quality as a possible quality metric for each aspect of quality.

State-of-the-art error taxonomies used to manually evaluate MT and HT have converged to very similar configurations. Typically, error classifications are hierarchical, with distinct top-level categories following Gideon Toury's principles of adequacy and acceptability (Toury, 1995). For example, the core variant of TAUS Dynamic Quality Framework (DQF)-MQM harmonised error typology used by the WMT-2020 has Accuracy and Fluency among the top-level categories (with further subcategorisation).

Typically, annotated translation solutions are categorised as:

- content errors that reflect the relations between source and target texts (adequacy and accuracy aspects of quality); and
- language errors that reflect the readability of the target text and can be assessed independently of the source text (fluency aspect of quality).

Error taxonomies designed for MT and HT may differ as to whether they include explicit pragmatic (fitness for communicative purpose, stylistic discrepancies) and macro-textual errors (e.g. cohesion, deixis, choice of generic nouns). In error analysis of HT the pragmatic component is either accounted for as a separate major category or is included in the content errors on the assumption that they are identified by comparing the source and target documents, given the purpose of translation.

When error analysis is applied to MT (especially when the experimental unit is a sentence), macro-pragmatic errors are impossible to account for as such. In sentence-level setup, at best, they account for stylistic and register errors as separate categories.

Stylistic errors can be treated as their own top-level category as in WMT datasets or grouped with the languages errors (Zaretskaya et al., 2016, see).

Despite according to the documentation, the MQM framework is designed with the functionalist approach in mind , this functionality was rarely used in real-life MT research projects before very recently (for example, Sanchez-Torron and Koehn (2016) used the Accuracy/Fluency dichotomy).

At the same time, the inbuilt ability of the MQM framework to dynamically setup a custom configuration and include a holistic component that can address functional and macrotextual aspects of quality (reader impression, sentiment, clarity, fitness-for-purpose, cohesion) is a step forward (or in the direction of HT notion of quality) compared to the earlier error systems for MT were based on a few microtextual error types such as missing word, extra word, wrong word order, incorrect word (Llitjós, Carbonell and Lavie, 2005; Vilar et al., 2006).

Another common feature of error annotation schemes applied to MT and HT is the implementation of some weighting scheme that recognises unequal severity of annotated errors. Within MQM "an error can be minor (if it doesn't lead to a loss of meaning and it doesn't confuse or mislead the user), major (if it changes the meaning) or critical (if it changes the meaning and carry any type of implication, or could be seen as offensive)" (Dugast et al., 2016).

Severities are used to weight errors according to the formula in Equation 4.3 (from Specia et al., 2020, p.745, who refered to Sanchez-Torron and Koehn (2016)):

$$\begin{gathered} MQMScore = 1 - \frac{1 \times n_{min} + 5 \times n_{maj} + 10 \times n_{cri}}{n} \\ \text{where } n_{min}, n_{maj}, n_{cri} \text{ are counts of respective severities,} \\ \text{and } n \text{ is the number of words in the sentence/document} \end{gathered} \tag{4.3}$$

In the area of HT annotation, severities are known to be less reliable attributes (Huertas-Barros, Vandepitte and Iglesias-Fernandez, 2018, p.33). However, no inter-annotator agreement is reported for benchmarks released by WMT.

According to organisers, annotation was crowd-sourced to human translators from *Unbabel* community. Based on the example of annotation, given in the Findings paper, the annotation was carried out on isolated sentences (Fonseca et al., 2019, p.3).

An error annotation setup similar to the one used to annotate students translations in our corpus, is described in Daems, Macken and Vandepitte (2013); Daems et al. (2017). Interestingly, they advocated a two-step approach to TQA even in the error annotation setup of "problems for adequacy and acceptability". The authors reported confusion with distinguishing the two aspects of quality at the initial stage of the annotation. Their proposal for alleviating that difficulty consisted in carrying out annotation for accuracy and acceptability separately and creating a different setup for each aspect (including and excluding ST, respectively). They also introduce the idea of severity weights to each error category that were adjustable taking into account the type of text and translation task.

Like other researchers in the field (Kübler, Mestivier and Pecman, 2018; Verplaetse and Lambrechts, 2019), Daems et al. (2017) relied on *brat* annotation tool (Stenetorp et al., 2012)[15]. Its major limitation is that it does not accept bitext formats as input. However, the alternative annotation tools (such as Appraise (Federmann, 2010)[16]) are even less suitable as they can be restricted as to the length of annotated items or are more cumbersome to install and get access to (*BLAST* (Stymne, 2011), *MT-EquAl* (Girardi et al., 2014)[17], *ACCOLÉ* (Esperança-Rodier, Brunet-Manquat and Eady, 2019)[18]).

The recent improvements in MT, which brought it to the verge of human parity, are likely to promote the use of error annotation to provide required fine-grained diagnostics of the output. Promising directions of analysis include the adaptation of error taxonomies to a given language pair and to the linguistic phenomena that are particularly challenging to neural MT (Klubička, Toral and Sánchez-Cartagena, 2018).

**Direct assessment: established practices** DA is a fairly recent quality type in QE research. MT quality metrics (i.e. evaluation methods) competition at WMT had used DA since 2016, and in 2019 it was borrowed as a learning target for *QE as a metric* task. QE captures general human judgments or perceived quality, as opposed to scores obtained from fine-grained annotations with errors or from post-editing. According to Fonseca et al. (2019), the introduction of human scores instead of PE effort shifts the QE application focus from professional translators to developers and general public as MT users. DA for

[15] brat.nlplab.org/
[16] github.com/cfedermann/Appraise
[17] github.com/hltfbk/MT-EQuAl
[18] lig-accole.imag.fr/app.php/login

MT metrics is still collected via crowd-sourcing on *Amazon's Mechanical Turk* (Akhbardeh et al., 2021, p.9), with additional checks for assessor consistency as described in Graham et al. (2015). The organisers report high correlation between DA and ranking results (Pearson's $r$ between 0.920 and 0.997 for various language pairs including out-of-English, p.174). However, we can expect a change of strategy in favour of professional translators as raters in line with recommendations by Läubli et al. (2020).

To be fair, QE datasets (including with DA scores) always include annotations by professional translators. According to Fomicheva, Sun, Fonseca, Zerva, Blain, Chaudhary, Guzmán, Lopatina, Specia and Martins (2020), each segment was annotated by two language service providers, each time by a team of three professional translators, resulting in six scores per segment. The correlation analysis between HTER and DA scores revealed diverging behaviour of those annotation methods (Pearson's $r$ did not exceed -0.344). This confirms that HTER and DA capture various dimensions of quality. As shown by Fomicheva, Sun, Fonseca, Zerva, Blain, Chaudhary, Guzmán, Lopatina, Specia and Martins (2020), a sentence with a critical error which is easy to post-edit would score low in DA, and low in HTER.

The annotation procedure followed the general approach by Graham et al. (2013), with modifications and quality checks suggested by Guzmán et al. (2019).

We describe key parameters of this assessment setup because we reproduced them in our own annotation experiment (see Section 4.4.3). Graham et al. (2013) proposed a continuous visual analogue 100-point scale, where the selection is done using a mouse-operated slider dispensing with traditional interval-level Likert scale presented as radio-buttons.

This solution is motivated by the continuous nature of translation quality, and does not force the raters to squeeze estimates into a fixed number of discrete Likert categories. The latter is believed to be a source of inconsistency in human judgments. Besides, it allows to understand the degree to which one translation is better than the other (similar to ranking), and it is better suited for some statistical methods.

The continuous DA implementations have an integrated quality control mechanism which allows to omit assessments from the less reliable annotators. It is based on testing whether each annotator produces the expected scores on the 20% of translations of "known quality". Those include (a) repeated samples ('ask_again') and (b) spam items

('bad_references'), i.e. translations is deliberately degraded in a way that still requires reading the sentence pair. This intra-rater agreement mechanism (and scores standardisation with z-transformation) are used to increase inter-rater reliability results.

To account for systematically different levels of leniency, i.e. legitimate disagreements between the raters, the individual scores are standardized by normalizing the raw values with respect to the mean and standard deviation of all scores assigned by a given individual judge (which is z-score conversion). Finally, the z-scores are averaged across all reliable raters to produce the final translation quality score.

Some aspects of evaluation protocols remain subject to controversy in the field. On the one hand, Graham et al. (2013, 2015) suggests that fluency should be judged independently of adequacy. They argue that it reduces the cognitive load and makes the assessment task better defined. Following this approach, Daems, Macken and Vandepitte (2013) implemented a two-step error annotation. They asked their participants, first, to annotate the targets for acceptability (using no other resources but the target text) and then, in a separate effort, to compare the source and target to annotate adequacy errors, following the descriptions of ranges for each quality aspect. Their experimental setup ensured that the two stages of the annotation were as independent of each other as possible.

On the other hand, there are claims that the two aspects might be indistinguishable, especially if both aspects are evaluated within the same setting in the same annotation effort. For example, Callison-Burch et al. (2007, p.140) stated that "people have a hard time separating these two aspects of translation" and that "the distinction might be false".

In some experimental setups the raters are specifically asked only about accuracy ("How accurately does the above candidate text convey the semantics of the source text?" (Hassan et al., 2018)). In the *FLORES* setup (used to annotate WMT-2020 data), the participants are asked to rate instances "according to the perceived translation quality" and the guidelines indicate that accuracy and fluency are not treated separately (e.g. "the 51–69 range represents a translation which is understandable and conveys the overall meaning of source string but contains typos or grammatical errors") (Guzmán et al., 2019, p.6109). Nonetheless, this setup implements the manual filtering procedure that evaluates translation quality and fluency for each potential datapoint separately and discards very disfluent translations.

They also described automatic checks to filter out faulty translations, i.e. sentences

that had:

- a lower perplexity than a threshold from a count-based n-gram language monolingual model;
- sentence-level char-BLEU score outside of a set range (two disparate or too similar output from two human translators);
- more than 33% of transliterations;
- 50% of words from the source sentence;
- more than 50% (or more than 5 in total) OOV words.

Manual quality control included independent re-evaluation for any sets where the discrepancies between raters were more than 30 points on the accuracy 100-point scale.

We are not aware of the experiments that used DA to evaluate entire documents, though document-level ranking was implemented for both adequacy and fluency conditions by Läubli, Sennrich and Volk (2018).

In the realm of human translation, DA method can be compared with the holistic assessment proposed for grading translation in various educational and professional settings.

### 4.3.2 Reliability and Validity

The outcomes of a research are only to be trusted if based on the reliable and valid estimates produced according to a sound methodology. This is especially true in the area of translation research, where evaluation procedures were shown to have a significant impact on findings and their interpretation (Läubli et al., 2020).

This subsection overviews practices that are currently in use to demonstrate the reliability of different annotation approaches labelling data.

In the area of human judgments about translation quality, especially in the area of MT, low inter-annotator agreement is a recurrent concern. It is explained, in part, by genuine disagreements between human experts about translation quality, attributed to differences in leniency levels, diverging interpretations of quality criteria and principled lack of a single, objectively correct translation for any given text.

**Adequate experimental setup** True intra- and inter-rater differences can be further confounded by the complexity of annotation setups. From this point of view, pairwise

ranking is characterised as the most straightforward and easy to explain, while post-editing and error annotation requires substantial professional and linguistic competence. DA was designed to collect perceived quality judgments from non-professionals, but it was also shown to be more reliable if professional raters are involved.

Apart from actual reliability scores obtained from one of the metrics, reliability of an assessment procedure can be assured by observing a number of conditions concerning the choice of raters, the presentation of each annotation task and their ordering to avoid known biases (e.g. reference bias).

Other necessary prerequisites of any annotation project include an unambiguous task (e.g. *Which text is better English?*) supported by written guidelines with examples. Some reasonable familiarity of the raters with the annotation environment and the scheme is also required. It is usually achieved through training and calibration sessions.

The claims of human-machine parity in translation have prompted a series of experiments aimed to re-visit the current practices in MT evaluation. Published recommendations include the following factors that were shown to improve the sensitivity of evaluation procedures to subtle differences between high-quality translations (Läubli et al., 2020):

- Choose professional translators as raters
- Evaluate documents, not sentences
- Evaluate fluency in addition to adequacy
- Do not heavily edit reference translations for fluency
- Use original source texts

Trustworthiness of an annotation experiment results can be demonstrated with regard to a number of aspects. Inter-rater agreement (IRA) scores reflect the extent to which raters' responses are concordant, i.e. how often the raters assign the same precise score, while Inter-rater reliability (IRR) measures the relative consistency of the responses, i.e. to what degree the scores assigned by raters are correlated, reflecting the same trend (Gisev, Bell and Chen, 2013).

The simplest form of agreement measurement is percent (or proportion) of agreed items. Correlation coefficients such as Spearman's rank correlation can be used to assess reliability as defined above. Most popular statistical measures, described below are used to report both agreement and reliability, with reliability being a more generic term.

The same metrics can be applied to responses by the same rater on multiple occasions to measure *intrA-rater* agreement and reliability. It can be used to disqualify raters whose answers are volatile on the same items.

Except IRR, some researchers report *validity* of the experimental results, i.e. the extent to which an assessment measures what it is designed to measure, or accuracy of a measure. Eyckmans, Anckaert and Segers (2009) compared several criterion-based and norm-based assessment methods for HT. They relied on Cronbach's $\alpha$ to assess reliability of the proposed *Calibration of Dichotomous Items-method*. Cronbach's $\alpha$ reflects validity of the method, i.e. internal consistency of a measurement scale of separate items (such as a cloze test) in the range from 0 to 1. This metric is based on variance and covariance statistics, which makes it inapplicable to nominal or ordinal scales.

**Reliability scores** Reliability for annotation tasks based on relative preference ranking (choosing the best candidate or best-to-worst ordering) or on a discrete rating scale (four, five- or seven-point Likert charts) is measured by the standard inter-rater agreement metrics.

Cohen's kappa is probably the most commonly used. A large-scale campaign to verify the human judgments in MT datasets revealed very low inter-rater agreement in the human evaluation tasks (as low as $k_{inter} = 0.37$, $k_{intra} = 0.52$ for WMT-2014 data (Graham et al., 2015)). In the same work it is suggested that Krippendorff's $\alpha$ might be a more appropriate metric for human evaluation tasks because, in addition to chance-correction and interchangeability of raters (also assumed in kappa), it allows for different magnitudes of disagreement and takes into account more than two coders (Artstein and Poesio, 2008). The values for both coefficients range between 0, signifying agreement at the chance level, to 1 for complete agreement.

Krippendorff's $\alpha$ coefficient was used to measure the inter-rater reliability by Yuan and Sharoff (2020), a research which is very relevant to the current project. For each aspect of translation, the authors reported alpha values over 0.74 for two independent annotators, who evaluated translations by scoring such components of quality as usefulness, terminology, idiomatic writing, target mechanics according to a grading rubric.

In case of DA continuous scale, reliability is ensured by elaborate statistics-based quality control measures that test internal consistency of crowd workers (intrA-rater

reliability) and filter out any participants who do not fit the stringent requirements. Additionally, as we have indicated above, the final scores are z-scored to smoothen the differences in rigour of assessment between otherwise consistent participants (Graham et al., 2013).

According to Chatzikoumi (2020) in the context of post-editing the "ideal procedure includes more than one judge and an inter-annotator agreement calculation". However, in practice there are only few studies that measure how often independent post-editors enter similar edits with regard to the same text span. It does not come as a surprise, because tracking and measuring disagreement about location of translation solutions that require editing is less straightforward than measuring categorisation disagreements. Besides, it seems difficult to measure the agreement on the edits introduced. Theoretically, these edits can display as much variability as any human translation.

Similar problems arise with the verification of error annotation schemes and environments, except in the latter case in addition to the agreement on error localisation, there should be agreement on error categorisation.

This makes error annotation is a multi-annotation task. It involves at least two sub-tasks: unit locating (unitising) and unit labelling (categorisation). There are few metrics that are designed to account for these two aspects, while the application of more familiar reliability measures requires reducing the task to only unitising or categorisation.

Lommel, Popović and Burchardt (2014, p.33) demonstrated two separate approaches to calculating pairwise Cohen's kappa for error annotation: the one that measures agreement on location (but not the nature) of annotated spans, representing each segment as a sequence of words, and the other that measures "whether the same issue classes were identified for each segment".

Their approach is criticized by Mathet, Widi̇ and Métivier (2015, p. 444-445), who suggested gamma ($\gamma$) inter-annotator reliability metric. This metric is used for verifying annotation in tasks such as named entity recognition or identification of text fragments associated with various propaganda techniques. Using this sophisticated metric imposes some limitations on the annotation setup. For example, embedded spans are not allowed and the annotation tool should support the unambiguous export of the annotation positions.

Alternatively, some studies demonstrate the consistency of quality estimates by reporting the correlation between final scores from independent raters. This approach does not put

to scrutiny the individual annotators' decisions but assesses the degree to which the outcomes of human evaluation are correlated (Eyckmans, Anckaert and Segers, 2009).

Disagreements are either resolved by an independent rater or discarded. There is a trend to retain opinions of internally and otherwise reliable annotators to show-case true disagreements on quality.

In cases where the disagreements can not be resolved by going down to the diverging categories or mismatching spans, or where the direct assessment methods are used the typical way to take into account all the available opinions is to use averages as the benchmark.

## 4.4 Quality Judgments for Student Translations

This research uses four types of quality labels and scores to identify the most reliable approach to quantify quality, at least, in terms of its automatic predictability, i.e. with regard to objective connection between the label/score and the linguistic form.

We use several approaches to represent quality and define datasets that are partially intersecting (see Figure 1.1 at page 9).

The quality labels/scores used in this project include:

- binary labels reflecting professional status of translators ('student', 'professional'): the biggest functionally-homogeneous class of English sources found in the RusLTC includes 334 English sources (with 3,134 multiple translations) that are annotated as informational, essays, educational, speech or interview and treated as a single 'mass media' register. To compare student translations to professional, one random student translation from each of ST set was retained. Professional category is represented by 404 published translations of English mass-media texts that were selected from a larger pull based on the functional similarity to the STs in the student collection.
- binary categorical labels ('bad', 'good'): several top and several bottom translations to each of the 57 English mass-media source texts offered to students as translation contests exam or routine assignments and assessed by translation teachers on the document level (183 translations in the 'good' category and 177 'bad' translations,

average document length is about 450 words).

- document-level and sentence-level continuous quality scores from error statistics: the final version of the RusLTC error-annotated subcorpus had 553 usable translations for 46 source texts (12,369 sentences, 7,166 of them had marked error spans).
- document-level and sentence-level continuous scores from a direct assessment experiment: to collect DA scores we set up an environment following recommendations for MT annotation suggested in Graham et al. (2013, 2015); Guzmán et al. (2019); Läubli et al. (2020). For this effort we selected a subset of 30 source texts that had binary labels and error annotations. Additionally, these translations were produced 'independently', i.e. in a competition, test or exam setting when no prior discussion of the source text was available. This subset was previously used to study typical source language (SL) difficulties in student translation (Kunilovskaya et al., 2022). We were able to obtain quality scores for 140 translations, counting 3,318 sentences (3,224 of them were matched with sentence-level error-annotated collection).

This section gives more details on annotation procedures to get or verify these labels/scores (excluding professional status) and on the associated inter-rater reliability studies.

### 4.4.1 Binary Quality Labels and their Verification

Our quality annotated corpus includes the total of 360 student translations to 57 English sources with binary document-level labels. These labels were obtained from 'grade' attribute in RusLTC metadata. Grades are real-life final assessment of students' translations produced to fulfill exam or routine course-work tasks in a number of universities and reflected in official records. A dozen of ST sets were translated at student translation competitions held by partner universities across a number of years.

All translations were assessed by translation experts (either university teachers of translation and/or professional translators). In all cases assessors were presented with the ST and a set of its translations (sometimes up to 40). Though each translation competition and each institution, where translations were graded, had their own quality requirements, they were not targeting a specific quality aspect or translation task. The grades represent overall perceived document-level quality assessed in a hybrid setting combining the properties of ranking and DA. In more than half of sets the final grade is

a consensus result for a competition jury or an exam board.

For the purposes of this research, we used only 1–3 top ranking translations, i.e. translations that received the lowest grade and 1–3 bottom translations, i.e. translations with the highest grade in the set. They were labelled ‘bad’ and ‘good’, respectively.

In an attempt to validate the integrity of this dataset and obtain estimates of human performance on it, we conducted a re-evaluation study. Three trained linguists with extensive experience in practical translation education were asked to compare two student translations to their ST and to each other. For each translation they indicated whether it was ‘good’, ‘bad’ or no difference in quality was perceived (the ‘same’ option). Figure 4.1 has a screenshot of one of the tasks implemented in *Google Forms*. The three documents in the task are accessible via links, they open in new browser windows.

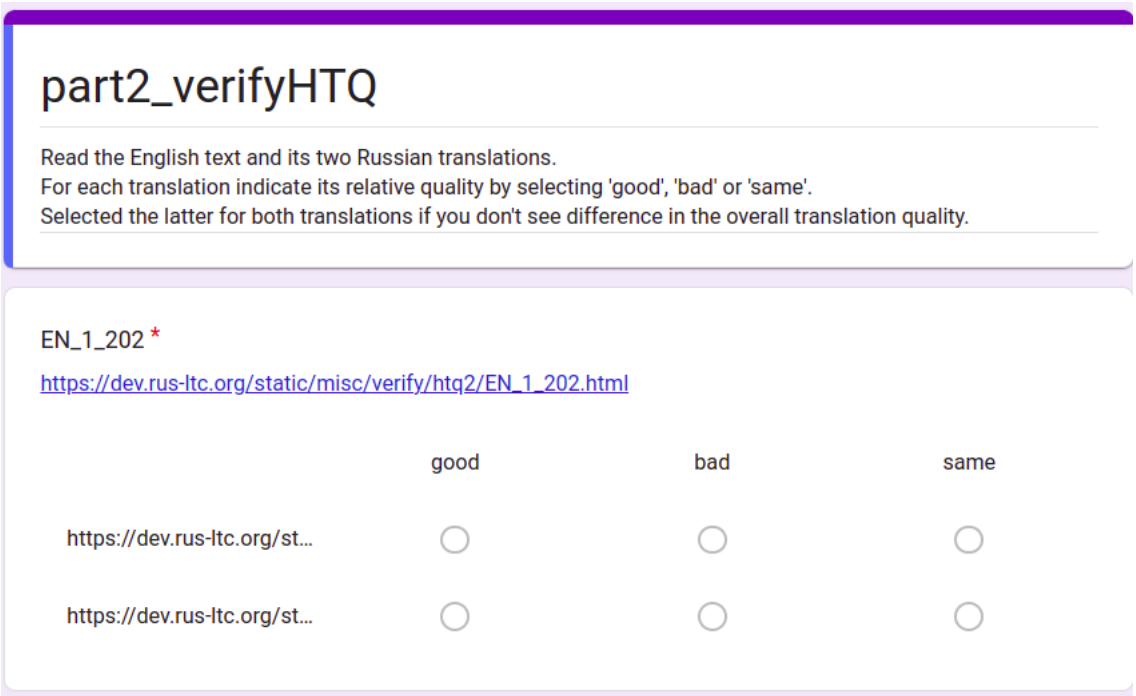



Figure 4.1: Annotation environment for binary labels verification

To create the tasks, we randomly selected 40 out of 57 ST. ‘Bad’ and ‘good’ translations were random documents from the respective quality categories for each ST. The average document size is 29.5 sentences, 535 words. The experiment was arranged in several blocks of tasks to avoid tiring the annotators with exhausting assignments.

In this setting, all three raters demonstrated total harmonic agreement in 21 sets out

of 40, i.e. 53% of cases. In three sets the raters returned three different responses (7.5%)

Inter-rater reliability for three raters across 80 decisions by each participant was computed as weighted Krippendorff's alpha ($\alpha$), interpreting the contrasts in translations quality as an interval scale, where the disagreements between 'good' and 'bad' had bigger penalty than between 'good' and 'same' (i.e. 'good': 2, 'same': 1, 'bad': 0). The overall reliability in the experiment in this setting was $\alpha = 0.52$.

If we look at the pairwise measurements in Figure 4.2, it becomes clear that one of the annotators (CoderN) had a tendency to agree with the other two participants less. They selected the 'same' category more often.

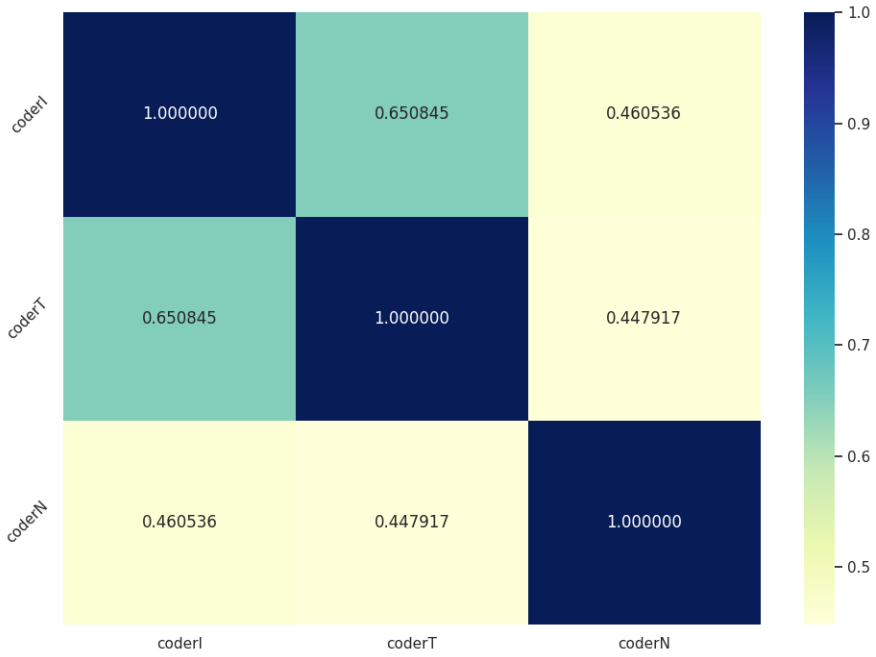



Figure 4.2: Pairwise Krippendorff's alpha for re-assessment of documents with binary labels

This indicated one of the limitations of that dataset: the 'good' label did not reflect absolute quality. In fact, these translations were the best in the set, but not without their flaws. In an educational setting, it is typical to have a relative, rather than an absolute grading scale.

To get an estimate of human performance on the re-annotated sample against the labels extracted from the *RusLTC* corpus, we had to discard five translation pairs (out of

40 sets) where two annotators could not decide on the quality (and used 'same' option), because the gold labels from *RusLTC* are binary and do not have 'same' category (two sets, four documents). We also had to remove three sets (six documents), where no consensus between the raters was achieved.

Given these edits to the original set, annotators verified the difference in translations quality with the accuracy of 91%, F1-score 0.914.

The total of 10 translated documents that were re-annotated as 'same' or caused disagreements between the raters were filtered out to get the quality-labelled dataset described in Section 3.2.1.

### 4.4.2 Error-based Quality Scores

This section gives a brief description of the error annotation setup underlying the error scores in our project. We discuss its reliability and implications for covering error statistics into continuous quality scores for sentences and documents.

***RusLTC* error annotation setup** The error annotation was setup following the general principles described above for evaluating human translations in an educational context. It is based on a pre-defined taxonomy which spans 30 error types arranged as a three-level hierarchy. The top level distinguishes between content and language errors. Annotators can select higher-level tags if a selected span does not fit any of the lower-level options (an alternative way to add flexibility to an error taxonomy is to have an 'other' category).

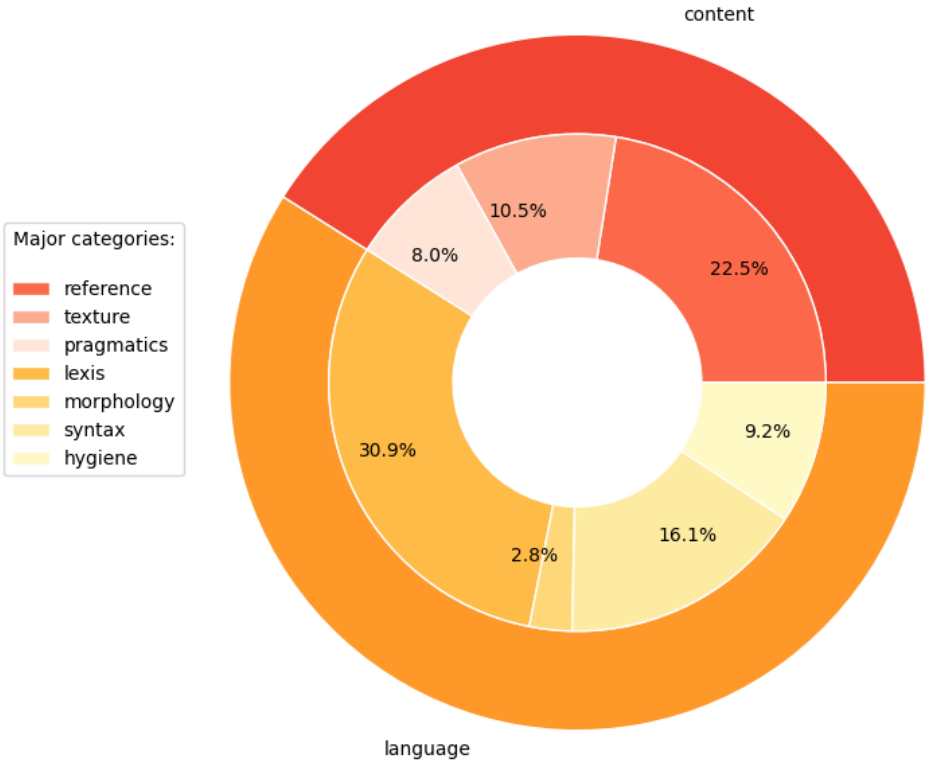



Figure 4.3: Percentages of major error categories to total errors

The distribution of errors across the major categories within content-language dichotomy is shown in Figure 4.3. In this collection content errors account for 41% of all annotated spans.

Additionally, annotators can assign *weights* and *technology* attributes to selected spans.

The former is a traditional three-member scale (*critical, major and minor*). Critical content errors are detrimental to the transfer of the original message. They significantly misrepresent the content of the ST, so that key information of any type is lost or can be misunderstood. Language errors are critical if they are obviously binary (Pym, 1992). They violate a basic TL rule and are immediately noticeable without being typos. There errors can be said to affect adequacy as they disrupt smooth flow of information and affect the perception of translation by undermining the trust in its faithfulness. Solutions annotated as minor errors (or not at all) include recommendations for possible improvements or non-binary errors.

The latter is a typically educational arrangement meant to describe annotated spans

as linked to professional competences (e.g. lack of background information, inability to suggest a more idiomatic solution, lack of language competence in either SL or TL, etc). Another educational feature of this taxonomy is positive evaluation tag, known as *kudo*, which is used to highlight good translation solutions for known translation problems, especially when most of other translators fail to find adequate renditions. Finally, annotators are encouraged to leave a note, which can be brought up by hovering a mouse over an error tag.

*RusLTC* error taxonomy and detailed statistics on tags in the error-annotated subcorpus used in this thesis are given in Appendix D. sec:mygold The annotation is performed online using *brat* annotation tool, a popular instrument for annotating text spans. The web-based interface was used as an educational tool to provide feedback to students in 2013-2018. All translations were anonymised before annotation. They were made available to students via a link (see a screenshot in Appendix D. Error statistics was used to rank student translations for a source text as a basis for relative evaluation of them.

| | 2013 | 2014 | 2015 | 2016 | 2017 | 2018 | total |
|---|---|---|---|---|---|---|---|
| sources (docs) | 13 | 10 | 9 | 6 | 5 | 3 | 46 |
| translations (docs) | 161 | 99 | 113 | 61 | 67 | 52 | 553 |
| translations (wc) | 60,691 | 43,917 | 41,972 | 27,949 | 26,850 | 24,418 | 225,797 |

Table 4.1: Translations by year of annotation

In Table 4.1 we report basic quantitative parameters of the error-annotated parallel subcorpus (based on raw text after sentence alignment). Its translated side counts over 225 thousand words across 553 multiple translations for 46 English STs. Like in all our parallel subcorpora, the sentence alignment was manually corrected; the sequence of sentences in each document reflects the true order of the published ST. The number of annotated error spans is 11,347 in total. On average an annotated document consists of 22.37 (+/-8.68) sentences (with a minimum of 10 and a maximum of 51). Each ST has from 3 to 26 annotated translations, 12.02 (+/-5.69) on average.

Figure 4.4 shows that the distribution of errors from major error categories was relatively stable across six years of annotation, except the first year, which included a

dozen of documents from earlier periods.

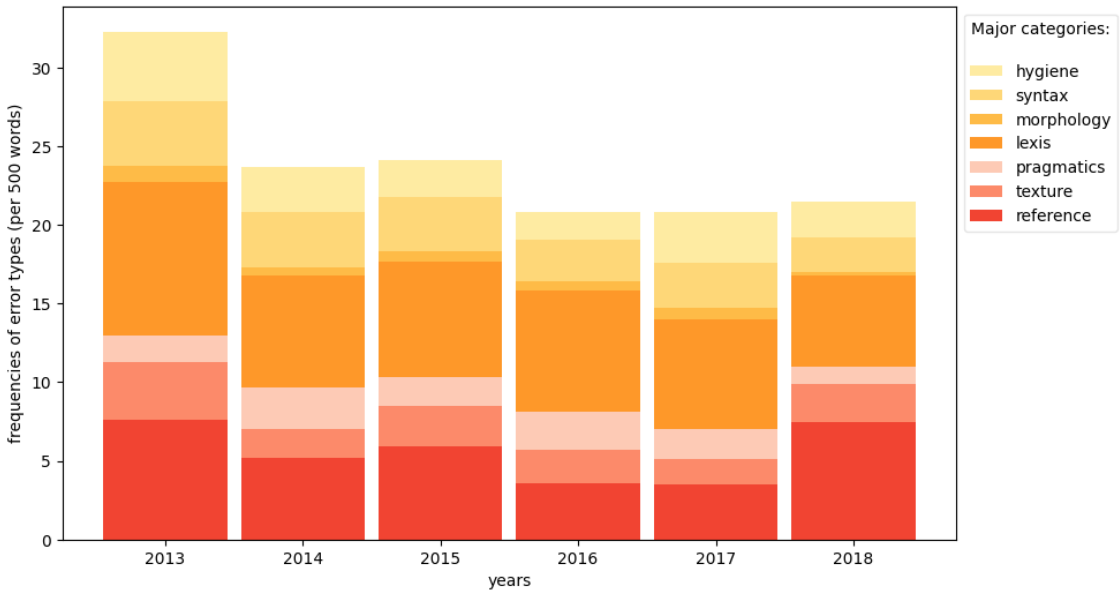



Figure 4.4: Distribution of error tags across years of annotation

**Reliability studies** The inter-rater reliability in the described annotation environment was measured as agreement between raters about the error location, and about the type and seriousness of errors in agreed spans in two separate studies. The experiments involved two and three trained linguists with experience in teaching practical translation. They worked independently and followed formal guidelines with exemplified description of the taxonomy (see the corpus website[19]).

The first study was designed to test the taxonomy and detect error categories that were most confusing or overlapping. The sample under analysis included 27 translations to 6 original newspaper texts (7,874 tokens in total). The number of translations to each ST varied from two to seven. The two raters selected the same span in translations in 64% of the average number of annotated errors (343 out of 539 cases). However, the total errors for the raters differed substantially (Rater 1: 630, Rater 2: 448), indicating the dissimilarities in tolerance to imperfections. Based on the agreed spans only, the raters agreed on the category of errors (collapsed to the major error categories) in 80.5% of cases reaching Krippendorff's $\alpha = 0.605$. We noticed that the raters had a tendency to agree on critical content errors, while the major areas of disagreements were related to the number

[19] rus-ltc.org/static/references/kunilovskaya_error_annotation_guidelines.pdf

and severity of language errors. The difference in rigour of analysis, especially with regard to adherence to the TL norms, was also reflected in the amount of positive evaluation, i.e. number of kudos. Rater 1 who was annotated more errors overall, also praised students twice less than the other rater.

The second study was based on a slightly improved error taxonomy and included the same two raters, and a colleague who was new to the proposed assessment system. They were asked to annotate 17 translations of one ST (571 tokens), and to award a final grade on a 20-point scale. The result show that the previous calibration session substantially increased the agreement between trained raters (including in the rigour of judgment). The result of the experiment were affected by disagreements with the new participant. The agreement on error location between three raters was 29%. For these spans (recognised as errors by three raters), the agreement about their category reach 76.76% ($\alpha = 0.535$).

With regard to the post-annotation overall document-level assessment expressed as a final grade on a 20-point scale, the agreement between the three raters was much higher: $\alpha = 0.734$ (measured with Krippendorff's alpha for interval scale).

To sum up our observations from these studies: the least reliable components of the proposed error-based assessment include (i) total number of errors, (ii) gravity of language errors, (iii) positive evaluation (kudos). Besides, we established that raters had a tendency to agree more on the quality of the low-ranking translations than on the high-ranking ones. For the bottom ten documents the agreement between three raters on the scale was $\alpha = 0.607$, while for the top ten translations $\alpha = 0.265$. The ratio of content to language errors was one of the most stable parameters of the annotation across all raters and experiments; it consistently was 40:60 proportion.

Further details of these reliability studies can be found in Kunilovskaya (2015).

Note that for the two cross-annotated samples we used annotations from the more lenient rater. The rest of the subcorpus had a single error annotation by one of the two translation teachers from the first experiment.

**Converting error statistics to quality scores** The error-annotated 553 translated documents include 12,369 sentences with 11,861 annotated errors (approximately 21.45 errors per document). Mean source sentence length was 21.87 words. More that half of the sentences (7,226) have at least one annotated span.

Generating quality scores from annotations was not a trivial matter, especially for the sentence level. We faced two problems.

First, unfortunately, our annotation was not originally designed for matching the target side spans with the respective parts in the source document and was not based on aligned sentences. The spans were not limited to word, phrase or even sentence boundaries. Therefore, we found it tricky to transfer annotations to documents that were (manually) sentence-aligned at a later date. The procedure included expanding the annotated span to at least 15 characters[20] to ensure a unique match between annotations and target side of the parallel corpus. This allowed us to generate sentence-level scores for aligned data. Due to this, in our setup document scores were not produced as aggregates of the sentence-level scores.

Admittedly, the procedure was not entirely reliable, but we managed to recover almost all annotated errors. Average ratio of true number of annotated errors per document to those found in sentence-level representation of the same document was 1.05 (+/-0.087). The Pearson correlation of overall quality scores calculated for documents directly and calculated as averaged sentence-level scores was 0.887.

Besides, we failed to find a reliable automatic way to link target side errors with their triggers on the source side, which limits the potential of converting this subcorpus to the word-level error-annotated format used by WMT. The community would definitely benefit from a tool that would accept sentence-aligned documents (bitexts) for error annotation. However, an attempt was made to overcome this difficulty in a manual cross-annotation setup. The results are reported in Kunilovskaya et al. (2022).

Second, it was not clear whether to use weights for various error categories and severity levels. If yes, how exactly to weights categories and severities and how to combine them.the

There seem to be two formulas in related research. Apart from the one adopted by WMT and given in Equation 4.3, there are recommendations from *TAUS*[21] that essentially leave the selection of weighting strategy to the researcher, but generally suggest (i) deciding on weights for error categories given the type of translation, and (ii) using the

[20] this threshold was established empirically

[21] www.taus.net/resources/blog/measuring-content-quality-with-error-typology-step-by-step-guide (Accessed: 23 July 2022)

following weights for severities and kudos (see Equation 4.4, where the score is presented as a percentage instead of a decimal fraction).

$$TAUSScore = 100 - (\frac{1 \times n_{min} + 10 \times n_{maj} + 100 \times n_{cri} - 1 \times n_{kudo}}{n}) \times 100 \tag{4.4}$$

where $n_{min}$, $n_{maj}$, $n_{cri}$ can be already weighted by error category weights

Similar to Formula 4.3, *TAUS* formula suggests to convert the calculated penalties to a positive evaluation of translation quality and normalise to the word count.

All scores in this study were normalised to the ST word count. ST word count was used as the normalisation base ($n$ in denominator) to reduce the impact of translation size on the calculation outcomes.

Arguably, the weighting strategy can introduce additional subjective bias to the final scores. Therefore, we created datasets with raw scores and weighted scores (marked with and asterisk throughout). **Raw (unweighted) scores** were generated as plain unweighted error counts of content errors for *accuracy*, language errors for *fluency*, and total errors for *overall quality score* (tq).

**Weighted scores** were calculated on an elaborate scheme which takes into account our observations from annotation reliability and the general domain informational text type. The *overall quality score* uses weights for severities, and it is increased if a translation solution was annotated as a good one (kudo, or positive assessment annotation). There were 677 kudo annotations in our collection.

**kudos** $\times 1$

**critical errors** $\times 5$

**major errors** $\times 2$

**minor errors** $\times 1$

Weighted *accuracy* and *fluency* scores were weighted for severities plus for individual error categories, following a scheme shown in Table D.3, Appendix D.

Figure 4.5 and 4.6 demonstrate that weighting improves the distribution of error scores, but might introduce yet another subjective factor in translation quality assessment.

Figure 4.6 shows that weighting also improves the spread of the data across the measuring scale which should be beneficial for machine learning experiments.

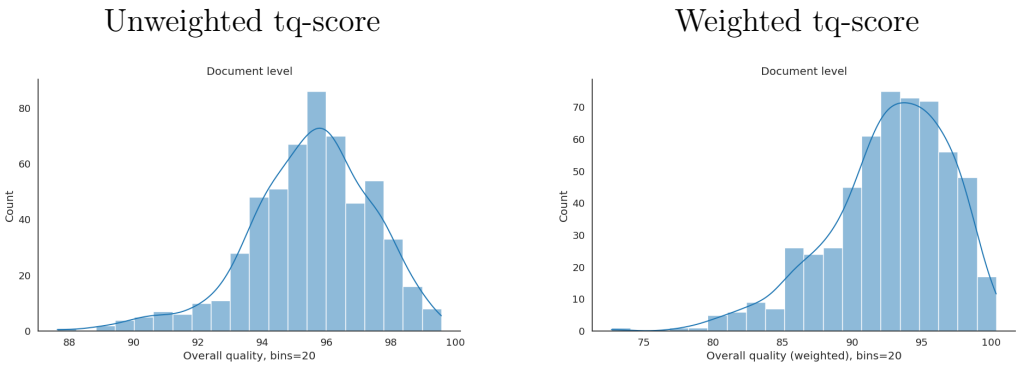



Figure 4.5: Distribution of overall error-based quality scores for 553 documents

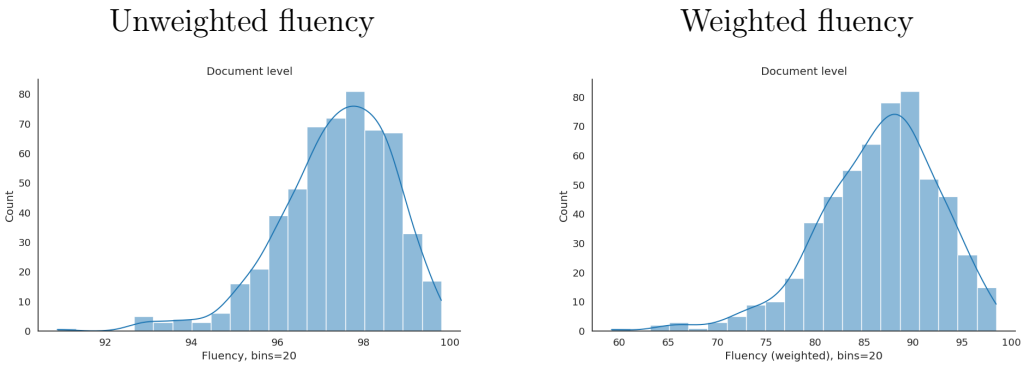



Figure 4.6: Document-level fluency scores for quality aspects

The same strategy to generating error-based quality scores was applied at sentence level. Out of 12369 sentences in the dataset, 7226 had some sort of annotation (including kudos). While producing sentence-level dataset we skipped 365 sentence pairs that were shorter than three words either on the ST or TT side of the parallel corpus.

Note that the sentence-level scores reflect issues related to the document texture (cohesion, coherence, logic, etc) if those were located in a given sentence. Care was taken to count errors annotated as discontinuous only once.

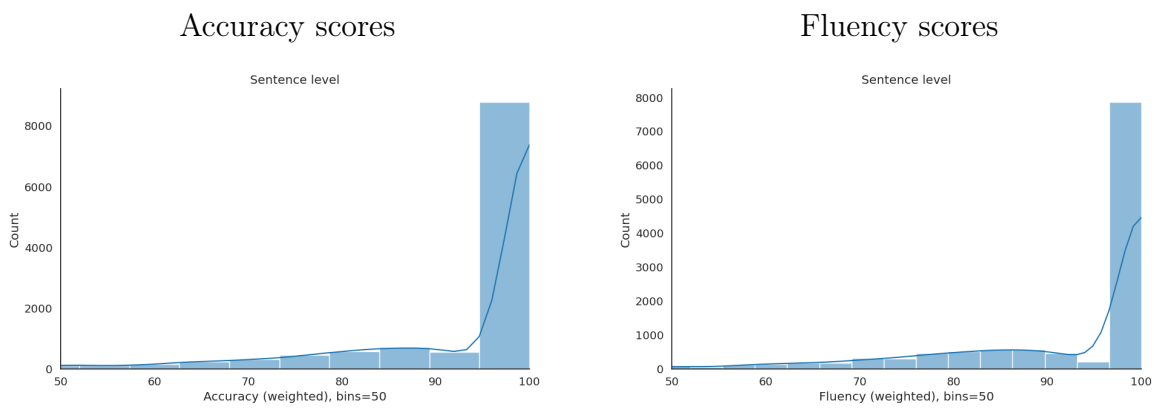



Figure 4.7: Sentence-level weighted scores for quality aspects: accuracy and fluency

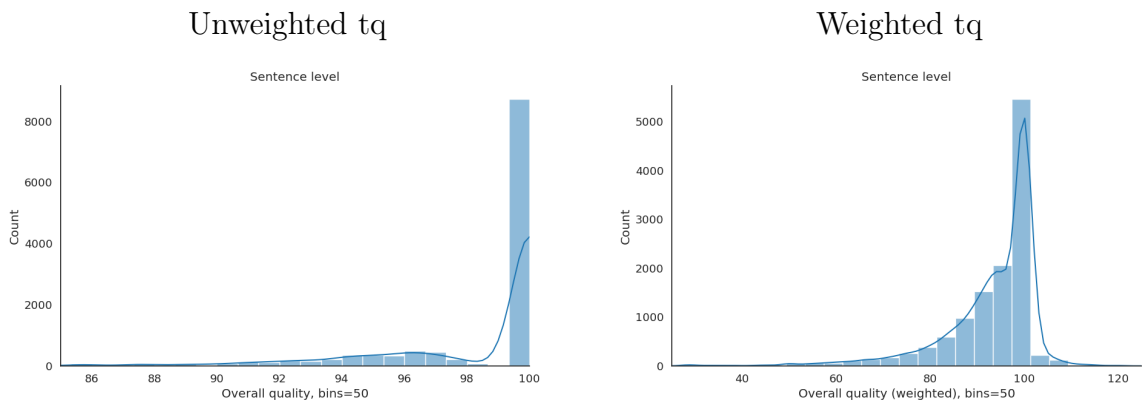



Figure 4.8: Sentence-level overall quality: unweighted and weighted

As can be seen from the histograms in Figures 4.7 and 4.8, sentence-level scores are heavily skewed to the right, which can be a problem for ML algorithms. The distributions of document-level accuracy, fluency and overall quality scores are closer to normality.

### 4.4.3 Direct Assessment

This section has a description of an annotation project, specifically designed to get quality labels from a controlled experimental setup which follows the requirements and recommendations discussed above.

**DA experimental setup and results** The Direct Annotation experiment was organised in October 2020 - January 2021. It involved 12 final-year bachelor degree students majoring in Linguistics or Translation Studies (Russian L1, English at least B2). The

participants volunteered to do the annotation as part of their internship; the arrangement was approved by the University authorities. The participants were also selected from a larger pool of volunteers following the recommendation their respective course leaders, who characterised them as best of their cohorts (based on academic achievements in their majors). The students' work was graded upon the completion of annotation tasks and a written report which contained cross-linguistic analysis of source and target texts with a commentary of translation errors. All participants were using aliases instead of real names throughout the experiment and in all communication of results and progress.

The experiment was set up on *QuestionPro*[22] website, a flexible environment to design online surveys. Following the most recent recommendations for organising quality annotation campaigns by Läubli et al. (2020), an annotation task was structured to include the entire document (or coherent extract from large texts) and to present annotators with the source and the target. The annotators were asked to move a slider on a 100-point scale to indicate perceived translation quality.

At the initial stage of the project, there were separate tasks for assessing fluency (in a reference-less monolingual setting) and accuracy (in a bilingual setting). An annotator never received the same document for both settings to avoid familiarity bias. However, the results on control items with 'known' quality indicated that annotators struggled with distinguishing the two aspects, and this distinction was abandoned for a syncretic quality judgment. The annotation task was formulated as follows: "Read the source text. Use the slider to indicate how much you agree that the text in bold is an adequate translation of the original English segment, given the context." Figure 4.9 has a screenshot of one of the assignments.

[22] www.questionpro.com/

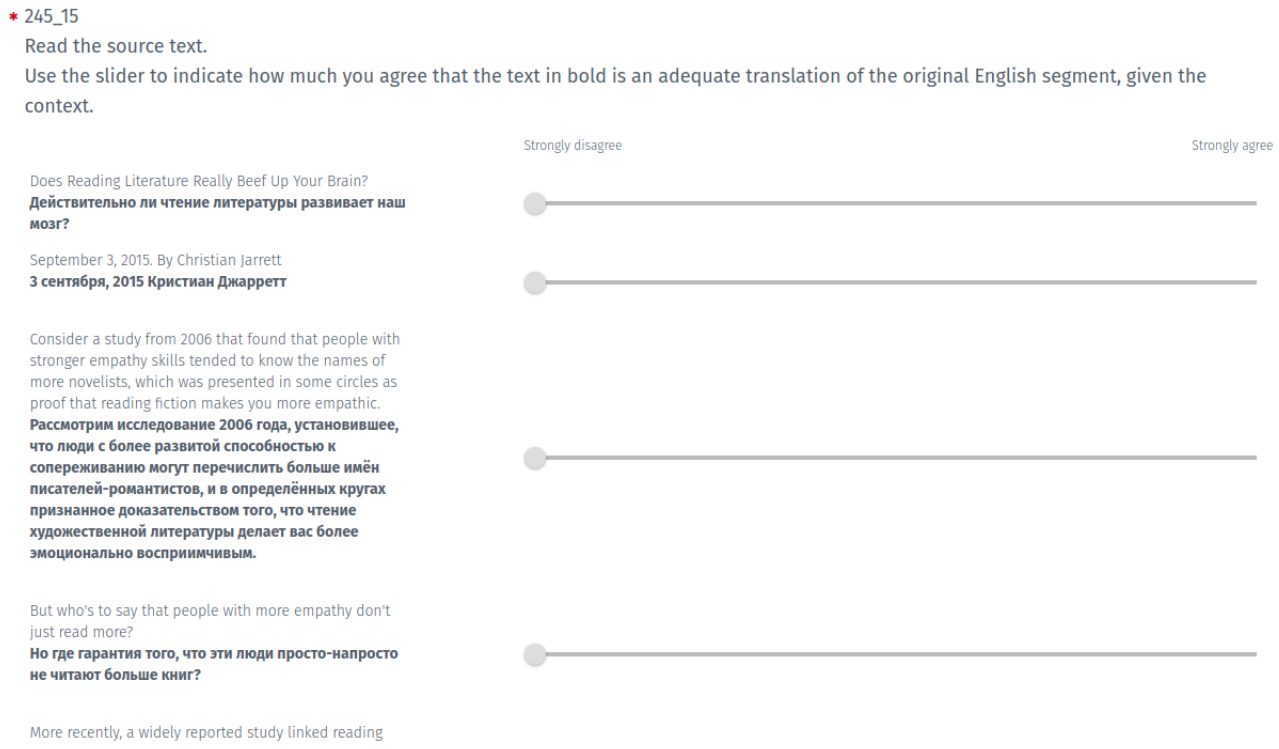



Figure 4.9: Part of a bilingual document and the annotation task

Before starting on their independent annotation tasks, the participants were explained the purposes of the experiment and had to complete a 30-minutes questionnaire which presented them with a number of translation quality assessment tasks and ensured that they understand the key concepts and terms (e.g. adequate translation, accuracy, fluency, source, target). Besides, they went through a calibration session to train the participants to use the annotation environment and harmonise the rigour of judgments, while using a slider rather than a Likert scale. The calibration results were discussed at a conference and distributed as a written report among the participants.

The independent work was arranged in four stages. At each stage annotators received a batch of 20 annotation tasks, one per document pair. Each complete annotation task (a parallel document of about 21 sentence pairs) was saved separately. The participants were given enough time to complete each batch and were encouraged not to do all 20 tasks in one go.

The documents for annotation were carefully selected from the intersection of the datasets with binary labels and quality scores from error annotation as shown in Figure 1.1 (page 9).

The dataset for annotation comprised 150 multiple translations to 30 English sources, 3,849 unique sentence pairs, including spam items and one repeat document for every participant. Spam items were documents with artificially reduced quality to test the integrity and sanity of annotators. Repeat items were exactly the same texts offered to each annotator two months later in a different batch to measure their internal (intra-rater) reliability.

Each source text had 2-3 good and 2-3 bad translations which were never offered in the same batch. In fact, one reason to involve many participants was to avoid presenting an annotator with alternative translations of one source text, at least within one batch, to avoid familiarity bias. To achieve this, we grouped annotators into two teams of six: three translators and three linguists on each team. Each team received their own set of 75 annotation tasks, i.e parallel documents (1,936 and 1,913 sentence pairs for Team 1 and Team 2, respectively).

To obtain trustworthy annotations, we identified most internally reliable annotators on each team who spotted the spam sentences and returned most consistent results on the repeat items presented to them in different batches. In one of the teams, only two people managed to overcome the established quality control thresholds. Therefore, we had to discard the lower-scoring rater in the other team. Those four raters were treated as interchangeable. This means that the entire dataset had double annotations, half from one team, half from the other team. The inter-rater agreement for the document-level annotation achieved weighted Krippendorff's alpha[23] of 0.541 (140 documents, 3,224 sentence-pairs).

For sentence-level, Krippendorff's alpha was 0.463 for z-transformed scores (across 3,224 instances).

Further on, for each team we highlighted individual sentences where both raters returned scores that had more than 30-points difference. Out of 3,318 true instances (sentence-pairs excluding spam and repeat items) in 3,090 cases the raters yielded scores that were less than 30 points apart. The 228 sentences with greater mismatch often were errors in alignment or encoding, and the raters treated them differently. In some translations by-line or date of the publication were omitted in translation, and this led to

[23]We treated 0-100-point scale as an interval scale and z-transformed values before IRR calculations as recommended in (Graham et al., 2013)

disagreements.

To get the final scores, we averaged judgments of two raters. Following the practice adopted at WMT, the dataset also contains the average of scores z-transformed by rater using Equation 4.5. These scores are referred to as *da_ mean* and *da_ zmean*, respectively.

$$z = \frac{x - \mu}{\sigma} \tag{4.5}$$

where $\mu$ is mean of the scores by rater,

and $\sigma$ is standard deviation of the respective set of scores

Sentence-level datasets with scores from error annotation and from DA experiment were combined to get one dataset with five primary scores (accuracy, fluency, tq, da_mean, da_zmean) for intersecting items to enable comparison of the quality scores generated by various methods.

There were 94 DA-annotated sentences missing from the error-annotated dataset, which was filtered for very short sentences (e.g. *Liquid Gold* – Вода по цене золота [Water at the price of gold] (RU_1_325_7_s8); *Right?* – Согласны? [Agree?] (RU_1_268_4_s45)). Those sentence pairs were removed from the sentence-level (but not from the document-level!) dataset to ensure consistency.

| | documents | sentences | tokens | |
|---|---|---|---|---|
| | | | EN | RU |
| error annotated | 553 | 12,369 | 250,916 | 225,782 |
| inc. with DA | 140 | 3,224 | 62,441 | 57,102 |

Table 4.2: Parallel subcorpora with error and DA annotation

Table 4.2 brings together the basic quantitative parameters of the dataset with error annotation and DA annotation.

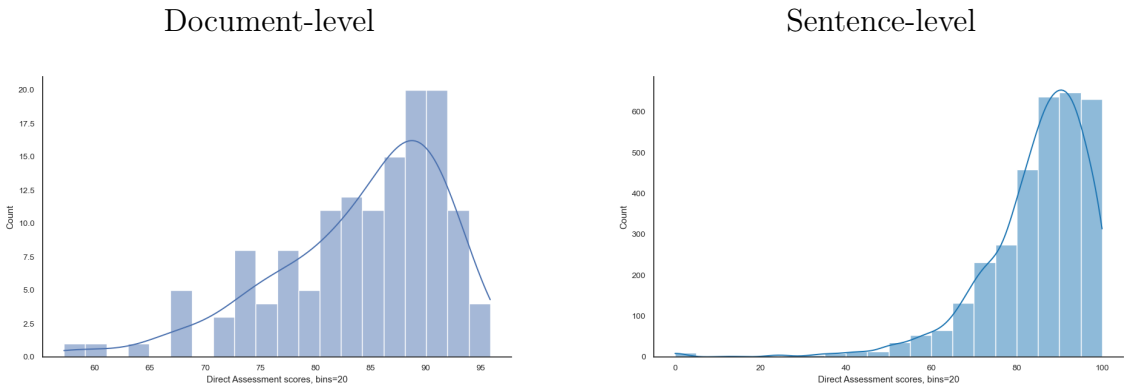



Figure 4.10: Distribution of scores from DA experiment for documents and sentences

We provide a visualisation for the distribution of DA scores at document- and sentence-level in Figure 4.10. Similar to scores from error annotation, sentence-level dataset has a more noticeable right skew: there are many more sentences with high-quality scores than with low-quality scores.

## 4.5 Summary

This is a key section in this thesis. It presents the theoretical concept of quality in translation and practical approaches to quality quantification. It is focused (i) on automatic quality estimation methods for machine and human translation (Sections 4.2.2 and 4.2.3 and (ii) on the best practices in producing manual quality benchmarks for NLP experiments (Section 4.3). The last part of the chapter has a description of annotation procedures used to obtain quality labels and score in this research.

In TS, quality is defined with the emphasis on extralinguistic context, including the purpose of translation, intended audience and medium (pragmatic aspect of quality, usually referred to as adequacy). MT research is mostly focused on accuracy and fluency aspects, where accuracy is viewed as (lexico-)semantic similarity between ST and TT, and fluency is approached as readability/grammaticality in the TL. The relative nature of quality was highlighted with regard to three factors, including the purpose of assessment, parameters of communicative situation with translation and comparison with alternative translation candidates.

Comparing approaches to quality in HT and MT, we have emphasised the limitations

of some MT methods due to the predominantly document-level nature of HT, greater variation of language and higher quality expectations attached to HT.

Section 4.2 characterises translation quality estimation as a computational linguistics and NLP task in MT, alternative to reference-based quality evaluation. It can be formulated at various language levels, with sentence level being more practicable for the current MT systems and language representation models. For anyone with the TS background, the idea to assess translation quality on isolated sentences is questionable.

This chapter offers an overview of the current best-performing approaches in MTQE and summarises the two projects that attempted to extend MT estimation methods to HT. We demonstrate that current state-of-the-art in MTQE relies on pre-trained contextualised embeddings fine-tuned on quality labelled data, and researchers' attempts are focused on explainability. Document-level MTQE task waits for further advances in technology to be successfully tackled and builds appropriate datasets meanwhile. We include a critical description of a popular MTQE feature set from *QuEst++* used as a baseline in this research.

Document-level HTQE based on a careful selection from a elaborate set of 360 types of hand-engineered surface features achieved Pearson's $r = 0.72$ outperforming *QuEst++* on the dataset proposed in Yuan's (2018) research. In their experiments feature-learning approach was inferior to explicit features at sentence level. However, a critical analysis of their learning setup (in particular, very repetitive translations for a very few STs) warrants doubts about robustness of this approach. The other research casts HTQE as an unsupervised sentence-level task based on learning similarity between ST and TT.

The last two sections of this chapter discuss the second component of any quality estimation task – quality labels. We present a fairly detailed analysis comparison of reliability and application of the major competing assessment methods in HT (rubrics and error annotation) and in MT (error annotation, direct assessment, post-editing). It has been repeatedly emphasised that when it comes to collecting human judgment about translations, it is best to resort to professional expertise, provide a rigorously formulated definition of quality for each text type/communicative situation and have a calibration strategy in place. The latest recommendations also require translations to be assessed in the true document order. In MT there is more awareness that these methods capture various dimensions of quality, and there is more pessimism about getting humans to agree

about quality. It is obvious that in MT there is a pressure to adjust quality benchmarking methods to the changes in MT quality delivered by the introduction of NMT. We tend to agree that the distinction between aspects of quality (accuracy and fluency) is even more difficult to capture in an annotation experiment.

While describing the best practices in producing quality benchmarks in MT we focus technical details, such as the choice of top-level error categories, use of severities, ways to convert error statistics to actual quality scores and annotation environments. A separate subsection discusses organisation of reliability studies for various annotation setup. We note that measuring IRR for error annotation or post-editing is particularly difficult because ideally it involves taking into account the agreement about the location of the annotated spans and their categorisation. Further caveats heeded in our own experiments include the recommendation to use Krippendorff' coefficient in IRR studies and to apply metrics to z-transformed (standardised) DA scores instead of absolute values.

It was also repeatedly shown that the outcomes of supervised ML experiments are contingent on the distribution of scores in the dataset.

Finally, we provided detailed descriptions of three annotation procedures used to obtain or verify various quality labels/scores in our datasets. The results were reported in terms of IRR and scores distribution at document and sentence levels. We demonstrated that in our various experiments the agreement between university translation teachers (nominal-scale assessment and error annotation) and final-year under-graduate students (DA experiment) as raters ranges from 0.467 to 0.734 in various settings, with the lowest result seen for sentence-level DA by students.

# Chapter 5 Translation Detection

This chapter establishes the premises of the investigated hypothesis that the amount of translationese manifested in a translation can be a good predictor of the level of professionalism and quality.

Sections 5.2 and 5.3 have the results of translationese classifications on professional and student translations to demonstrate that the selected hand-engineered features are, in fact, effective to capture translationese. To make sure that the proposed features are competitive in a translation detection task, their performance is contrasted with other representations including content-dependent features. If our hypothesis stands, we can expect that the more salient translationese indicators will also be good quality predictors. To shed more light on feature subsets and individual variables that can be considered most effective in capturing translationese, we analyse outcomes of automatic feature selection and perform univariate analysis for the three proposed types of features in Section 5.3.

The chapter opens with a description of the experimental setup for classification tasks including details on learning algorithms, generation of alternative representations and numeric data preprocessing (Section 5.1).

## 5.1 Experimental Setup

**Dataset and proposed translationese indicators** The input to classifiers was arranged as a table of 3055 rows by 167 columns. In that table each row represented a document (including sentence-aligned source texts for parallel subcorpora). Columns stored three labels, six meta descriptors of each document-instance (document ID, language, word count, number of sentences, raw text and lemmatised text) and 158 hand-engineered numeric features from the attempted manual feature sets.

The labels reflect three binary categories, listed below to introduce their shorthand names used throughout all experiments:

**stu/pro vs ref:** categories in a translationese classification based on either (random) student or professional translations;

**stu vs pro:** categories in a classification to capture professionalism (see Section 6.1.1 in Chapter 6);

**bad vs good:** categorial document-level quality labels from holistic assessment of student translations verified in a separate annotation effort (results on this dataset appear in Section 6.1.2).

In tabled results below linguistically-motivated translationese-aware feature sets are represented as follows:

**UD:** 60 features extracted from Universal Dependencies annotations (see Section 2.3.2);

**collgram:** 24 features representing abstract collocational properties of texts (see Section 2.3.3);

**ngram:** 10 features from 1-2-3-gram frequency lists and from n-gram-based language models (see Section 2.3.4);

**all:** combination of the above feature sets (94 features).

Adding actual raw and lemmatised texts into the table and setting the random seed helped to ensure that the same instances were found in the same order in the same folds regardless of the representation used: manual features or vectors from pre-trained embedding models.

**Alternative document representations** To demonstrate the competitiveness of hand-engineered features, we compare them with a number of alternative representations.

The chance-level baseline is calculated using a stratified strategy, which ignores the input features altogether. Its predictions are based on randomly assigned class value to each feature in a row respecting the class distribution in the training data.

The hand-engineered features are all relatively content-independent. We compare their performance to surface-feature-based tf-idf representation and to text vectors generated by mean-pooling of embeddings from several Transformer models. The vectorisation setup in each case was implemented as follows. *Tf-idf* is a numeric representation for documents which uses weighted frequencies of a corpus vocabulary items as vector components. Those vectors were calculated using a dedicated *scikit-learn* utility. We used all available documents in the Russian language as the underlying corpus. To reduce vocabulary size and increase informativity of this representation, the corpus was lemmatised. All proper names and their sequences were replaced with a *PROPN* tag (e.g *говорит Максимилиан*

*Чуперски* [says Maksymilian Czuperski] –> *говорить PROPN*). All digits were replaced with *X* (e.g. *12 октября 1982; 2,5; 1980-е –> XX October XXXX; X.X; XXXXs*). The vocabulary was built to include word unigrams, bigrams and trigrams. The vector size was limited to 5000 most frequent corpus vocabulary items. Other hyper-parameters were set to the default values from the library.

We also used four pre-trained contextualised embedding models to infer vectors for documents in our experiments. These models aim to generate distributed representations of document meaning (in a broad sense) and are content-dependent. The models are open-source and freely available from the *HuggingFace* repository.

We selected the models that were trained to solve various NLP tasks deemed more suitable for our purposes. The list below gives a brief description of each model.

**stsb-xlm-r-m:** a multilingual model[1] trained using Sentence-BERT (SBERT), a state-of-the-art (SOTA) framework to compute sentence representation (Reimers and Gurevych, 2019). SBERT is a generic name for contextualised embedding models fine-tuned on various sentence-level tasks such as question-answering, sentiment, paraphrasing, etc. The model we used was a multilingual XLM-R fine-tuned on STS task;

**TQmono-m:** a cross-lingual model for many languages[2] trained by the more successful architecture from *TransQuest* package (Ranasinghe, Orasan and Mitkov, 2020), to solve MT quality estimation task based on DA scores;

**ruRoberta:** a dedicated Russian model[3] trained as a full-mask language model using *RoBERTa*, an optimised BERT transformers training method proposed by (Liu et al., 2019);

**mdeberta3:** a multilingual version of *DeBERTa* (He, Gao and Chen, 2021), an improvement on *BERT* and *RoBERTa* word embedding architectures, trained on 2.5T CommonCrawl corpus for 100 languages [4]. At the time of writing, it is SOTA in word-level representation learning.

The first two models generate sentence embeddings, while the last two output word

[1] huggingface.co/sentence-transformers/stsb-xlm-r-multilingual
[2] TransQuest/monotransquest-da-multilingual
[3] huggingface.co/sberbank-ai/ruRoberta-large
[4] huggingface.co/microsoft/mdeberta-v3-base

embeddings, which were averaged to get sentence embeddings. Further on, the sentence embeddings were mean-pooled to obtain document vectors. Note that in most experiments we might not have enough datapoints to train these representations further on the labels available in our tasks in a fine-tuning setup.

For all experiments and representations, the input data was transformed with *scikit-learn Standard Scaler*. This function standardises the values of each feature column-wise (i.e. independently) to have zero mean and unit variance of 1. It helps to put all features on the same scale regardless of the original measurement unit, and avoid the situation when features with large variance dominate the learning process overshadowing the differences on other features.

**Algorithms** Both classification and regression experiments (reported in Chapter 6) are based on SVM algorithm as implemented in *scikit-learn*, a Python library for machine learning. This algorithm was shown to be superior to tested alternatives for translationese classification in a number of research projects (see, for example, (Ilisei et al., 2010)). It is also known to be robust to high-dimensional datasets where the number of instances is similar or even smaller than the number of features.

This research is focused on testing the hypothesis whether translationese indicators are good predictors of translation quality, and on exploring the relations between a number of diverse representations and quality labels. Our aims did not include achieving the best performance possible for given datasets, therefore, we do not report results for various SVM kernels and other parameters of algorithms that can be discovered through optimisation studies. Instead, the classification experiments are based on a linear kernel SVM and regularisation parameter $C = 1.0$ (default for SVM *scikit-learn* setting).

SVM uses pairwise distances between the datapoints from given categories and identifies instances with most similar vectors (called *support vectors*). Then, it calculates a maximum hyperplane separating these support vectors, by assigning positive or negative weights to features to put instances on the required side of the hyperplane and reduce algorithm error. Instances that are found on the opposite sides of the hyperplane are assigned to the respective categories. $C$ parameter is the cost of misclassification, or tolerance for errors. Higher $C$ values lead to fewer errors in the training set, and to a more overfit model, which would not generalise well. It will result in poorer performance on a test set.

Although the datasets in our experiments are reasonably well-balanced we kept the *class_weight* parameter of the algorithm set to *balanced*.

As a sanity check, we implemented a simple sequential neural model to replicate the performance of SVM-based classifiers on exactly the same inputs. The neural model was designed using *Keras* application programming interface (API) integrated into *TensorFlow 2.0* (tf.keras). It is promoted as the standard interface for deep learning.[5]

The neural model used document vectors as input and was compiled with two dense (i.e. fully-connected) hidden layers. The first layer was made of 16 neurons and was activated with Rectified Linear Units (ReLU) function, which output zero for negative input values and returned the input directly if the input was positive.

The output layer for the classification task was activated with a *sigmoid* function, which mapped all inputs in the real number domain into the range of 0 to 1. It transformed input values that were much larger than 1.0 to 1.0, values much smaller than 0.0 were snapped to 0.0. In fact, this layer output the probability of a test instance belonging to a positive class. The threshold of 0.5 was used to obtain the model predictions of binary categories.

The model was compiled with *binary cross-entropy* as loss function for the classification tasks. *Adam optimiser* was used for gradient descent calculations; each experiment was set to run for 10 epochs with batch size 8.

We also monitored the learning process with an *EarlyStopping* class. It was set to keep track of accuracy across 10 training epochs for each run of the experiment. The training stopped if the gain in the algorithm performance was less than 0.0001 for three consecutive epochs.

Another standard ML setting which prevents over-fitting, especially with high-dimensional data, is cross-validation. The data was split 10 times into a training and test sets in a cycle, respecting the balance of categories, so that each instance appeared in a test set at least once. We report the averaged results (and their standard deviation) across 10 runs (folds) of each experiment. All scores are expressed as percentages.

We use the standard evaluation metrics for classification problems as described in Section 4.2.1 (page 94).

[5]Both *TensorFlow* (`github.com/tensorflow/tensorflow`) and *Keras* (`keras.io/`) are open-source libraries for deep learning with Python.

**Feature Analysis Methods** Detection of the best translationese indicators is performed using recursive feature elimination (RFE) technique. We believe that some features might form functional subsets that would perform better together. Cross-validated RFE (RFECV) method was used to retrieve a dynamic set of most informative features for each dataset. RFECV applies an algorithm (a linear SVM in our case) to a feature space reduced by one feature with the least importance for each subsequent run of the experiment. The importance of features is established using weights assigned by the algorithm. The process is continued to obtain the best performance across several cross-validation folds (five, in our case). We scaled each features before selection to have zero mean and variance of 1.

We also used weights (and their visualisations) assigned by linear-kernel SVM in univariate feature analysis. According to explanations in Volansky, Ordan and Wintner (2015), high feature weights are a good indication of a significant difference between the two classes, but lower feature weights might not mean lack of discriminative power because feature weights can be reduced taking into account possible inter-dependencies among features in the feature set.

Apart from providing information about the more salient linguistic properties of each text category, feature selection allows to reduce the feature space by removing noisy features (if any), which might lead to an increase in the algorithms performance.

To establish major trends in translational behaviours that could be associated with the observed properties of each translation categories, we developed a sequence of statistical univariate tests that describe the differences between each pair of categories involved in a translationese study (source texts, translations and non-translated TL reference).

Non-parametric statistical tests were selected in all cases, because the values of most features were not normally distributed for either of the document categories. For example, here were only 12 features in non-translated subcorpus (out of 94 total) where the null hypothesis of normal distribution was not rejected based on *Shapiro-Wilk test*, a statistic test of normality. For many manual features, the compared categories of documents also had unequal variances. *Bartlett statistic*, which tests for homogeneity of variances across two samples, revealed that there were 68 feature with unequal variances in the translationese classification experiment based on professional translations, for example.

These peculiarities in the distributional properties of features also explain the choice

of *RFE* algorithm over *ANOVA* for feature selection. The latter compares between-categories and within-categories variation, and assumes normal distribution and equal variances between samples.

Statistical significance was tested using *two-tailed Mann-Whitney U Test for independent samples* and *Wilcoxon signed-rank test for paired samples* as a non-parametric alternative for data that violates the assumption of normal distribution made by Student's t-test.

Additionally, we measured the strength of the relationship between each variable (feature) and target labels using 5-fold cross-validated *Logistic regression* which returned macro F1-scores in the classification setup. The choice of the method is explained as follows. In many of our experiments we had to estimate the association between continuous and binary categorical variables, and could not use either *Cohen's D* effect size metric or regular correlation coefficients, because it is impossible to calculate the covariance in this setup. The use of a classifier as an association metric builds on an assumption that if we can accurately predict a categorical variable from a continuous variable, they can be considered correlated.

We also used scatter plots and line plots to present the results of PCA for some of our representations. By analysing covariances between features, PCA attempts to reduce multidimensional data to a few new features that aggregate maximum information of the original representation. Although dimensionality reduction simplifies the relations between datapoints, it can be useful to visually study how well each representation captures the distinction between categories.

All examples, unless indicated otherwise are taken from student translations. In the tables, we highlighted the best results for manually engineered feature sets and boxed the best results for content-dependent classifiers. The best results are defined as the results that were not significantly outperformed by any other model in that category at the confidence level of 0.05.

## 5.2 Translationese Classifications

This section presents the results of translationese classifications on professional and student translations. The quality of these classifications indicates how well a given feature set or representation can distinguish translations from non-translations. Hu and Kübler (2021)

raises a question about the level of accuracy that is high enough to accept a feature or a feature set as discriminative for the categories at hand. To give an idea of what performance is achievable, we report results from the best-performing content-dependent classifier in our experimental setup, namely, *mdeberta3*.

Statistical tests showed that there were no statistically significant differences between the performance on *mdeberta3* vectors, on *ruRoberta-large* and on *tf-idf* representation. For considerations of space, we omit the results on these models from the tables below.

Tables 5.1 and 5.2 aggregate results for a linear SVM classifier on three distinct hand-crafted delexicalised feature sets and their concatenation (*all*) for professional and students translations, respectively. They are put into perspective of chance level and contextualised embeddings results. Note that most of translationese indicators proposed by Volansky, Ordan and Wintner (2015) (and used as a baseline in some translationese research) were implemented in this study and included in either *UD*, *collgram* or *ngram* feature sets.

Results for a neural classifier (page 145), which was run in the same settings as an alternative for a linear SVM, appear in Appendix E. Systematic comparison between the SVM and neural classifier showed that out of 40 settings across various feature sets in four binary classifications, statistically significant differences between the algorithms were observed in only four cases (as measured with a two-sided t-test using F1-scores from 10 cross-validation folds, at confidence level $\alpha = 0.05$). It confirms that in terms of learnability, we achieved a reasonable performance level.

Generally, both translation varieties were easily separable from originally-authored texts on all attempted representations. Both lowest and highest F1-scores were seen on student translations: 65.63% on *collgram* feature set and 98.36% on *mdeberta3* vectors.

Structural morphosyntactic classifiers (UD) with linear SVM kernel demonstrated competitive performance and returned macro F1-score of 90.22% and 88.96% for professional and student translations, respectively.

| | Accuracy | Precision | Recall | F1 |
|---|---|---|---|---|
| chance | 47.72 (+/-5.93) | 47.72 (+/-5.93) | 47.70 (+/-6.00) | 47.58 (+/-5.95) |
| UD | 90.34 (+/-2.73) | 90.51 (+/-2.59) | 90.32 (+/-2.98) | 90.22 (+/-2.81) |
| collgram | 70.82 (+/-5.11) | 71.15 (+/-5.09) | 71.16 (+/-5.07) | 70.71 (+/-5.08) |
| ngram | 86.12 (+/-2.96) | 86.07 (+/-3.04) | 86.30 (+/-3.07) | 86.04 (+/-2.99) |
| all | 92.45 (+/-2.72) | 92.41 (+/-2.73) | 92.53 (+/-2.81) | **92.38** (+/-2.76) |
| mdeberta3 | 96.67 (+/-1.65) | 96.74 (+/-1.63) | 96.64 (+/-1.64) | 96.63 (+/-1.66) |

Table 5.1: Professionals: translationese classification. Support: N(pro)=404, N(ref)=497

The superior performance of content-dependent classifiers on this task is expected, because translations and non-translations can be on different topics. Classifiers that exploit lexical (string) features are likely to capture less interesting topical distinctions between the corpora.

| representation | Accuracy | Precision | Recall | F1 |
|---|---|---|---|---|
| chance | 50.05 (+/-3.72) | 48.65 (+/-3.80) | 48.62 (+/-3.88) | 48.59 (+/-3.84) |
| UD | 89.41 (+/-4.25) | 89.20 (+/-4.28) | 88.93 (+/-4.57) | 88.96 (+/-4.46) |
| collgram | 67.50 (+/-6.58) | 66.64 (+/-6.93) | 65.54 (+/-6.22) | 65.63 (+/-6.40) |
| ngram | 86.40 (+/-3.29) | 86.30 (+/-3.18) | 85.27 (+/-3.99) | 85.61 (+/-3.70) |
| all | 92.30 (+/-2.03) | 92.25 (+/-2.20) | 91.81 (+/-2.18) | **91.96** (+/-2.10) |
| mdeberta3 | 98.44 (+/-1.43) | 98.64 (+/-1.33) | 98.16 (+/-1.64) | 98.36 (+/-1.49) |

Table 5.2: Students: results of translationese classification. Support: N(stu)=334, N(ref)=497

For both translation varieties, structural UD features, which were focused in this study, significantly outperformed the other two manually-constructed feature sets. The results on 10 features from n-gram language models were 3-4% lower, while collocational features were 19-23% worse.

*Collgram* feature set, designed to capture collocational peculiarities of translational

language, returned the lowest performance (still much above the stratified random baseline). It also demonstrated higher volatility across the folds, with standard deviations as high as 5 and 6%. The performance of SVM classifier on UD features was more stable, especially for professional translations, where STD averaged at around 2.8.

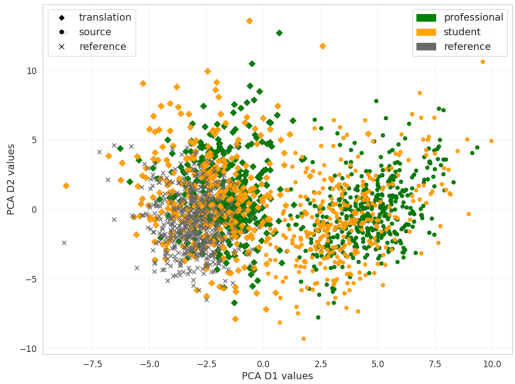
*Collgram* and *n-gram* features did not yield any real gain in classifier performance, when combined with UD feature set. Although F1-score on *all* features was a 2% higher for both classifications, the difference was not statistically significant (at $\alpha = 0.05$ confidence level).

**Performance of translationese classifiers on two translation varieties** Overall, the quality of translationese classifications for professional and student translations was approximately the same in most setups. This observation runs contrary to our expectations. We expected student translations to be less fluent and less similar to non-translations and, therefore, easier to detect. It is even more surprising that professional translations should return slightly higher classification results than student translations, particularly on UD features (F1-score 90.22% vs 88.96% for professionals and students by SVM, 90.43% vs 90.1% by neural classifier).

Although it is not a common practice, Demšar (2006) argues that non-parametric *Mann-Whitney U rank test* (aka Wilcoxon rank sum test) on two independent samples can be used to compare performance of one algorithm over multiple datasets. According to this test, there was no evidence that professional translations were more distinct from non-translations than students (at 0.05 confidence level) for either SVM or neural classifier on any representation (except *stsb-xlm-r-m* on both SVM and neural classifier). Based on the difference observed on *stsb-xlm-r-m*, we can conclude that professional translations were semantically less similar to the reference subcorpus than texts translated by students.

To visually explore the relations between text categories in the two experiments, we cast multidimensional document vectors into a 2D space using PCA. The scatter plot in the left-hand panel of Figure 5.1 presents non-translations, student and professional translations along with their sources using the first two principal components, constructed by PCA from UD vectors.

2D representation of document types




Distribution along PCA D1

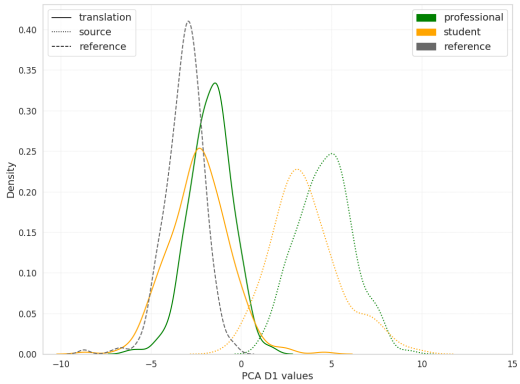



Figure 5.1: Visualisations of PCA transform based on *UD* features

Each marker represents a document from the respective colour-coded category. Coloured diamonds are used for translations, dots for source texts in English and dark grey crosses are non-translations in the TL. The most obvious distinction is between the two languages: Russian is on the left, English is on the right. It can be seen that all translations are shifted away from non-translations to the left towards source texts. The translation varieties are not clearly separated from non-translations and from each other. It means that the signal captured by fairly successful translationese classifications is distributed among the UD features and cannot be easily summarised by just two dimensions. To be fair, PCA offers a very crude simplification for the underlying datasets. PCA D1 shown in Figure 5.1 explains only 21% of the total multivariate variability in the full UD dataset, with further 13% captured on D2.

Importantly, we can see a difference in location between English source texts for professional and student translations: green dots of professional translation sources in left-hand panel in Figure 5.1 seem to be further away to the right from the centre of the plot. ST is a known factor in shaping the properties of translations, and this difference should be taken into consideration in feature analysis.

It can be argued that student translations (shown in orange) are the least compact cloud without a distinct location, while professional translations are a bit more homogeneous. This observation is supported by a more focused representation of document values on x-axis (PCA D1) in the right-hand panel using a kernel density estimation plot (a

smoothened and scaled version of a histogram plot). The group of three line plots on the left shows the variation in distributions of PCA D1 values across the document classes in Russian. The shape of the flatter orange solid-line curve for student translations suggest more variability in the data. Both student and professional translations are shifted to the right from non-translations (grey dashed line) towards the area of English sources. We can tentatively conclude that D1 captures language-contrast-related properties of documents.

By way of comparison, consider the same types of visualisations for our best performing representation, namely for *mdeberta3*, in Figure 5.2.

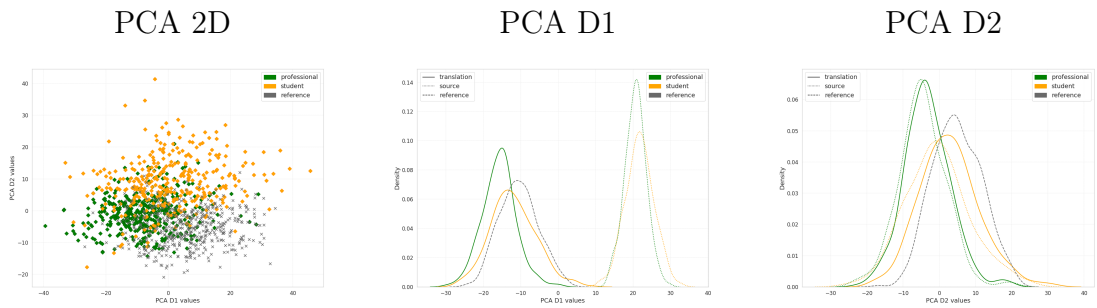



Figure 5.2: Visualisations of PCA transform based on *mdeberta3* representation

If we exclude English sources from PCA analysis (they overshadow all other distinctions on multilingual vectors), the three classes of documents in the TL are more or less distinct and are located in different parts of the 2D scatter plot. D1 (x-axis) seems to capture the differences between professional translations and non-translations, while D2 (y-axis) values separate student translations from the other two categories. Unlike the previous representation, translations (regardless of professionalism level) were not grouped together against non-translations. If anything, student translations appear to have more in common with non-translations than with professional translations. It means that *mdeberta3* did not really capture the commonalities of translations as an ontological text category. It is likely that the co-varying components of *mdeberta3* vectors, aggregated by PCA, reflect semantic (topical) differences between categories along each of the axes. The right-most panel in Figure 5.2 suggests that student translation source texts are indeed semantically more similar to Russian non-translations than English source texts of professional translations.

This impression is reproduced on *stsb-xlm-r-m* vectors[6] (see left-most panel in Figure 5.3). However, vectors from *TQmono-m*, a Sentence Transformer fine-tuned on MT quality estimation task, when squeezed into one component by PCA, revealed the commonality between the two translation collections. Both demonstrate a shift away from non-translations (in either language). Interestingly, student translations are located further away from non-translated reference than professional translations, capturing the anticipated difference in the amount of translationese (see central panel in Figure 5.3). Finally, the right-most scatter plot in Figure 5.3 shows that a dedicated Russian contextualised word embedding model also detects translationese per se (and not the confounding topical differences between text categories): it separates all translations from non-translations, but this distinction is captured on a weaker D2 component (y-axis), while D1 probably captures the variation in topical content.

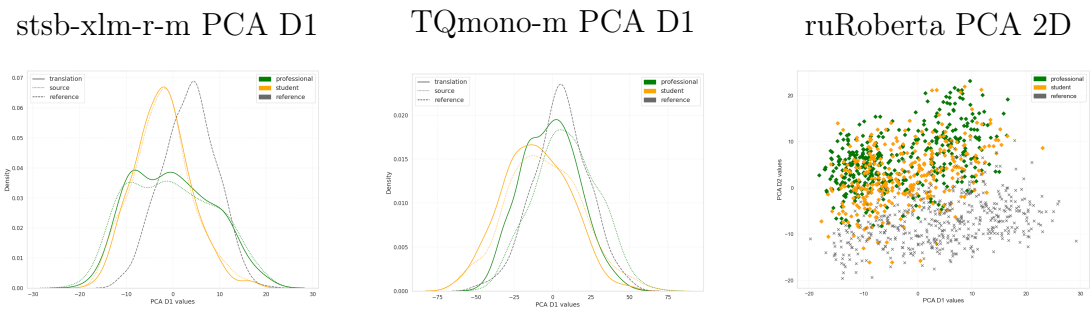



Figure 5.3: More informative aspects of PCA analysis on other contextualised embeddings

Despite contextualised embeddings seem to capture different aspects of documents relatedness, they achieved the same classification result: there were no significant differences between *TQmono-m, ruRoberta and mdeberta3*, with *stsb-xlm-r-m* being inferior only to *ruRoberta* and *mdeberta3* on professional translations by either SVM or neural classifier. On (more challenging) student translations mean-pooled word embeddings (*ruRoberta* and *mdeberta3*) significantly outperformed both sentence embedding models (*TQmono-m* and *stsb-xlm-r-m*) at least by SVM.

The next section describes UD features from the point of view of their usefulness for translation detection. We report the results of feature selection, based on SVM feature

[6] *Stsb-xlm-r-m* is a cross-lingual language model for many languages specifically trained to capture semantic similarities between sentences.

weights and carry out univariate comparisons between text categories to reveal most important translationese indicators and trends that they contribute to. We are going to rely on this analysis in describing the results of professionalism- and quality-related classifications with regard to our main hypothesis that translationese indicators can be related to translation quality.

## 5.3 Feature Analysis

This section aims to detect strong translationese predictors among the manually engineered features and provide a linguistic description of translationese in English-to-Russian language pair. The primary method to achieve this goal is feature selection, supported by univariate feature analysis and statistical significance tests.

Eventually, we want to see whether strong translationese indicators are correlated with professionalism and quality labels/scores and could predict them in a ML setup.

### 5.3.1 Best of UD features

**Feature selection** As discussed above, RFE was preferred over ANOVA for feature selection because many UD features did not meet the ANOVA requirement of normal distribution and equal variances across compared text categories. In contrast, RFE relies on internal SVM feature weights and gradually prunes less useful ones to reduce the feature set to required N. To infer N experimentally instead of using an arbitrary number of selected features, RFE can be run in a cross-validation setup (RFECV). In our experiments, the algorithm iterated N in the range [3 : size of the feature set], performed 5-fold cross-validation on each N and returned N with the highest mean F1-score across all folds.

These experiments returned $N = 33$ and $N = 42$ for translationese classifications on professional and student translations, respectively. After discarding noisy and collinear features, the performance of both classification improved. It went up by 1.3 percentage points from 90.22 to 92.04% for professional translations, and by 2.1 percentage points from 88.96 to 91.06% for student translations (based on macro F1 scores).

To have a closer look at the commonalities between translation varieties and feature

importance in these selections, the lists were intersected and ranked based on the weights assigned by the linear kernel SVM algorithm. Feature weights were obtained from classifiers run on full datasets (without reserving a fraction for testing). Experiments showed that the feature weights were approximately the same if the weights were averaged across the models learnt on each of the folds in a cross-validation setup.

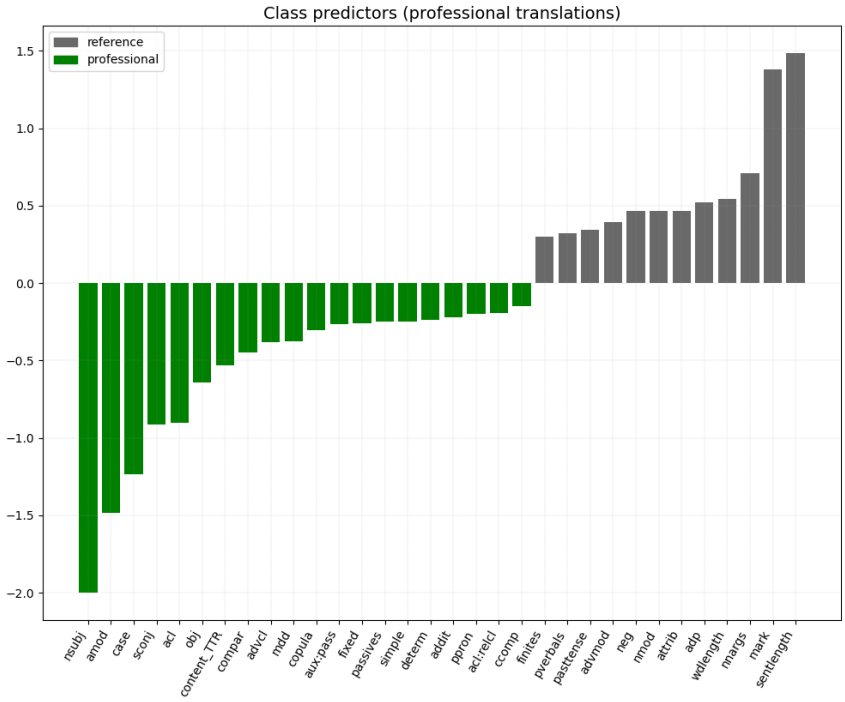



Figure 5.4: Weights loading the decision towards each category for 32 RFECV-selected items

Figures 5.4 and 5.5 visualise the weights assigned to the respective subsets of best features in each classification. Features with larger weights can be considered more predictive of the distinctions between categories.

Detailed explanations of shorthand used on the x-axis can be found in Appendix A.

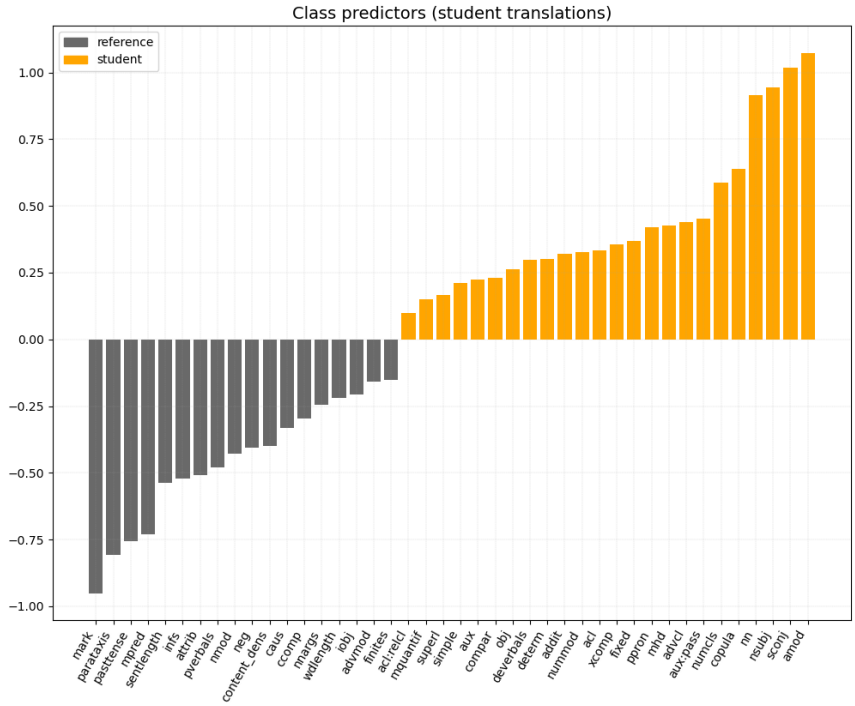



Figure 5.5: Weights loading the decision towards each category for 42 RFECV-selected items

There were 28 UD features shared between the optimal sets of 33 and 42 translationese indicators in professional and student translationese classifications, respectively. They can be considered strong predictors that capture translational specificity regardless the level of expertise.

The full list of features includes: *amod, sconj, mark, nsubj, pasttense, copula, sentlength, attrib, pverbals, aux:pass, advcl, nmod, ppron, neg, fixed, acl, addit, determ, deverbals, ccomp, obj, nnargs, compar, wdlength, simple, advmod, finites, acl:relcl.*[7] These features are shown as either 'translation predictors' (coloured bars in Figures 5.4 and 5.5) or 'non-translation predictors' (grey bars). Although the weights are shown as loading feature values towards a specific category (reference or translations), it is not clear how this choice is made by the algorithm and this aspect is ignored in our analysis.

For consideration of space, below we present these 28 features in typological groups, where possible.

[7]Because the weights for a feature could be quite diverging in professional and student classifications (Spearman correlation was 0.47), the features appear on this list in the descending order according to the absolute weights in the student classification, a hypothetical marked quality class in this research.

Two-thirds of the features listed above reflected various aspects of sentence structure, rather than frequencies of morphological units. Translationese-related syntactic phenomena were:

**nsubj, obj:** distribution of core verbal arguments (nominal subjects and objects),

**mark:** frequency of function words that identify another word as a head of a subordinate clause,

**acl:relcl, advcl, acl, ccomp:** frequencies of relative and adverbial clauses, finite and non-finite clauses that modify a nominal or function as clausal complements,

**finites, pverbals:** closely related to the previous features, the distributions of finite and non-finite verbal forms in adverbial function, including Russian-specific verb form referred to as a converb (*деепричастие*), a type of participle in the adverbial function,

**deverbals:** ratio of de-verbal nouns, naming of processes, actions, states, to the number of verbs,

**amod, attrib, nmod:** modifiers of noun, including adjectives, participles and other nouns in attributive function (e.g. genitive complement: *времена (финансовой) стабильности* [times of (financial) stability], *список грехов* [list of sins]),

**advmod:** non-clausal adverbial modifier (e.g. *но и для выражения сочувствия* [but also for offering condolences], *N существует в виде статичных изображений* [N exists in static images]),

**simple:** proportion of simple sentences to all sentences in a document,

**sentlength:** average number of words in a sentence, care was taken to exclude pseudo-sentences that had only numbers and punctuation marks, or contained less than two tokens,

**copula:** function words linking subject to a non-verbal predicate,

**neg:** negative particles or main sentence negation,

**addit:** additive conjunctions of parenthetical type (such as *what's more, in other words, furthermore, in this regard*), which are optional elements, external to the sentence structure.

Morphological classes, categories and forms, whose quantitative parameters were useful for translation detection, included:

**nnargs:** ratio of nouns or proper names as verb arguments to total of respective functions,

**ppron:** personal pronouns, excluding possessive pronouns,

**sconj:** subordinative conjunctions,

**aux:pass:** analytical passive verb forms

**pasttense:** verbs in past tense,

**fixed:** fixed multi-word expressions (MWE), mostly functional words (e.g. *то есть* [that is]; *до сих пор* [until now]),

**determ:** pronominal determiners (e.g. *каждый* [every], *этот* [this]),

**compar:** comparative degree of comparison for adjectives and adverbs,

**wdlength:** average number of characters in content words only to avoid co-linearity with features counting various function words.

There were five features that were specific for professional (but not student) translation classification: *adp, passives, mdd, case, content_TTR.*

To achieve the best result in predicting student translations, SVM considered distributions of 14 more features that were not selected in the classification based on professional translations: *numcls, parataxis, mpred, xcomp, aux, content_dens, superl, nummod, mhd, iobj, nn, mquantif, infs, caus.*

Features that were not selected by RFECV for either professional-vs-non-translation or student-vs-non-translations classifications included 13 items. They can be considered less useful for translation detection either due to the lack of considerable distinctions between the categories or due to the collinearity with other features. This list had values for the following features:

**tempseq:** cumulative frequency of temporal and sequential connectives (e.g. *короче говоря* [to put it briefly], *в конце концов* [finally, in the end]),

**discourse:** a dependency relation for interjections and other discourse particles (e.g. *в первую очередь* [in the first place], *в частности* [in particular]),

**epist:** epistemic markers of author's stance (e.g. *навряд ли* [unlikely], *можно не сомневаться* [no doubt], *вероятно, наверняка* [probably, definitely]),

**possp:** possessive pronouns,

**cconj, conj:** frequencies of coordinating conjunctions and respective syntactic relation

between two conjuncts,

**advers:** adversative (contrastive) discourse markers (e.g. *впрочем* [however], *несмотря на, тем не менее* [despite, nevertheless])

lexical density, counted as ratio of PoS disambiguated content word types to all tokens,

simple sentences,

**flat, compound:** two types of MWE: exocentric semi-fixed sequences such as proper names and dates (e.g. *Синдзо Абэ; 24 декабря* [Shinzo Abe, 24 December]) and endocentric compound MWE, usually hyphenated in Russian (e.g. *черно-белый* [black and white]),

**amod:** non-clausal adjectives and participles functioning as attributes (e.g. *финансовый центр* [financial centre]),

**obl:** frequency of non-core verbal arguments expressed by a noun (e.g. indirect objects),

**appos:** punctuated or parenthetical modifiers of noun (e.g. *мы, демократы* [we, the democrats], *мисс Пигги, персонаж Маппет-шоу* [miss Piggy, a Mappet show character]),

**interrog:** proportion of sentences ending in question mark,

**anysome:** indefinite and total pronouns (e.g. *кому-нибудь, чем-то, когда-то* [anyone, something, at some point]),

All features on the list above had very weak association with the labels in the range of 0.46 to 0.36, with the exception of oblique nominal objects (*obl*), a non-core verbal argument introduced in Russian by a preposition and, therefore collinear with the frequency of prepositions. Some features on this list seem to disprove a common anticipation by the translation textbooks of deviant use of respective items in translation. However, their use in 'translationese' contexts might have been subsumed by other more natural contexts.

The next paragraph explores the distributional properties of the best translationese indicators revealed in the feature selection process across document categories (translations, sources and non-translations). This contrastive-comparative univariate analysis helps to group features according to specific trends in translational behaviour reflected by the document properties.

**Univariate analyses** To understand more about the prominent patterns of language use in translations picked up by SVM classifiers, we performed a series of statistical tests on individual features identified as strong translationese indicators in the previous paragraph. It is aimed at revealing the impact of the language pair on the linguistic properties of translations. It is known that the distance between the languages is not the only factor that affects translational behaviour, but we do not have the necessary setup to measure the possible contribution from those factors.

For each feature, we compared its distributions in sources, targets and non-translations. The differences were tested for statistical significance using *Mann-Whitney U test for independent samples* and *Wilcoxon signed-rank test* for paired samples of source and target texts.

First, we identified the features that had significant differences between translations and non-translations (*tgt* and *ref* categories): these are translationese indicators by definition. Then, we established whether there were statistically significant differences between the two languages (*src* and *ref* categories) with respect to a given feature, and which language had greater values. Finally, for each feature, we performed pairwise comparisons between frequencies in translations and in each of the languages.

The outcome of the above sequence of tests can be viewed as capturing the amount of effort invested by a translator to reconcile differences between the two languages (if any) for a given feature. The features, representing various trends in translational behaviour, can be arranged in three major groups (see Table 5.3), depending on the adopted translation strategy and the scale of its implementation.

| | group | trend | description |
|---|---|---|---|
| | reproducing ST | anglicisation (ref<src<tgt) | TT values greater than in SL and much greater than in TL, possibly the use of SL patterns even when unprompted |
| | | shining-through ref<tgt=src | No difference between ST and TT values, both are higher than in TL |
| | | overuse of SL ref<tgt<src | Many translation decisions are prompted by ST, which has higher values than in TL |

— amount of effort ——— low

| | | |
|---|---|---|
| counteracting ST | underuse of TL<br>tgt<=src<ref | Lack of effort to add typical TL items when not prompted by ST |
| | normalisation<br>src<=tgt<ref | Insufficient efforts to bring ST frequencies in line with TL norm |
| | russification<br>src<ref<tgt<br>tgt<ref<src | Active use of TL patterns unseen in ST or effective counteraction of ST influence leading to overuse of TL items |
| | adaptation<br>tgt=ref<src<br>src<ref=tgt | No translationese: significant differences between the two languages are reconciled in favour of the TL norm |
| third code | SL/TL-independent<br>tgt<ref=src<br>ref=src<tgt | Translationese features with significant differences from both languages and no language gap |

Table 5.3: Patterns in translational behaviour

Besides, there was a group of features that were useless for a translationese study, because they had the same frequencies in translations and non-translations, and also did not distinguish the languages. In the classification based on professional translations this group included frequencies of adpositions (prepositions) and discourse particles such as interjections. In student translation data there were no statistically significant differences between document categories for interrogative sentences and adverbial quantifiers (e.g. *совершенно* [entirely], *чрезмерно* [utterly]).

Further on, we established the strength of association between each feature values and binary categorical labels (translation, non-translation) by measuring the quality of single-feature Logistic Regression classifier using macro F1-score (LogR). The results of significance tests and variable-label association analysis are collected in Table 5.4. The features selected by RFECV are sorted by LogR F1-score averaged for students and professionals. Statistically significant results are indicated by *less-than* and *greater-than* signs, the lack thereof is shown with an *equals* sign.

With regard to the strength of association between feature values and class labels (translations, non-translations), only half of shared features from RFECV selections yield F1-scores above the chance level (0.5) when used in a single-feature classifier (shown in bold in Table 5.4).

| | Professional | | | Student | | |
|---|---|---|---|---|---|---|
| feature | LogR | trend | means | LogR | trend | means |
| **nsubj** | **0.82** | overuse of SL | ref<tgt<src | **0.71** | shining | ref<tgt=src |
| **mark** | **0.74** | overuse of SL | ref<tgt<src | **0.68** | overuse of SL | ref<tgt<src |
| **acl:relcl** | **0.69** | overuse of SL | ref<tgt<src | **0.66** | overuse of SL | ref<tgt<src |
| **advcl** | **0.7** | overuse of SL | ref<tgt<src | **0.63** | overuse of SL | ref<tgt<src |
| **obj** | **0.72** | overuse of SL | ref<tgt<src | **0.61** | overuse of SL | ref<tgt<src |
| **fixed** | **0.56** | russification | src<ref<tgt | **0.58** | russification | src<ref<tgt |
| sentlength | **0.61** | NA | ref<tgt | **0.57** | NA | ref<tgt |
| **acl** | **0.62** | shining | ref<src=tgt | **0.56** | shining | ref<src=tgt |
| **nnargs** | **0.68** | normalisation | src<tgt<ref | **0.55** | normalisation | src<tgt<ref |
| **amod** | **0.51** | russification | src<ref<tgt | **0.53** | russification | src<ref<tgt |
| **attrib** | **0.51** | russification | src<ref<tgt | **0.52** | russification | src<ref<tgt |
| **copula** | **0.52** | overuse of SL | ref<tgt<src | **0.51** | overuse of SL | ref<tgt<src |
| **ccomp** | **0.63** | overuse of SL | ref<tgt<src | **0.5** | overuse of SL | ref<tgt<src |
| **nmod** | 0.36 | adaptation | src<ref=tgt | **0.5** | russification | src<ref<tgt |
| deverbals | 0.45 | normalisation | src<tgt<ref | 0.45 | russification | src<ref<tgt |
| addit | 0.4 | shining | ref<src=tgt | 0.43 | SL/TL-indep | ref=src<tgt |
| wdlength | 0.4 | NA | tgt<ref | 0.43 | NA | ref<tgt |
| pasttense | **0.51** | SL/TL-indep | tgt<ref=src | 0.42 | SL/TL-indep | tgt<ref=src |
| neg | 0.41 | normalisation | src<tgt<ref | 0.4 | normalisation | src<tgt<ref |
| advmod | **0.53** | overuse of SL | ref<tgt<src | 0.38 | SL/TL-indep | tgt<ref=src |
| aux:pass | 0.36 | russification | tgt<ref<src | 0.38 | adaptation | tgt=ref<src |
| compar | 0.38 | overuse of SL | ref<tgt<src | 0.38 | overuse of SL | ref<tgt<src |
| finites | 0.36 | adaptation | tgt=ref<src | 0.38 | overuse of SL | ref<tgt<src |
| determ | 0.36 | overuse of SL | ref<tgt<src | 0.37 | overuse of SL | ref<tgt<src |
| ppron | 0.36 | overuse of SL | ref<tgt<src | 0.37 | overuse of SL | ref<tgt<src |

| | | | | | | |
|---|---|---|---|---|---|---|
| pverbals | 0.36 | adaptation | src<ref=tgt | 0.37 | adaptation | src<ref=tgt |
| sconj | 0.36 | russification | src<ref<tgt | 0.37 | russification | src<ref<tgt |
| simple | **0.67** | underuse of TL | tgt<src<ref | 0.37 | underuse of TL | tgt<ref<src |

Table 5.4: Properties of 28 UD features identified as discriminative between translations and non-translations by RFECV

Table 5.4 shows that the more salient deviations from non-translation for both translation varieties can be described by the same trends.

Half of translationese indicators in both translation collections had frequencies, which were shifted towards SL value and were higher than expected in the TL. In these cases translators seem to reproduce patterns observed in the SL (overuse of SL patterns, shining-through or anglicisation). For example, the overuse of SL patterns was seen for *nsubj, mark, acl:relcl, advcl, obj.*

A different translational trend is associated with features, where translators actively counteract the influence of the ST and bring the frequencies in translations closer to the TL norm. Underuse of TL items, normalisation, russification and adaptation together explain 13 (professionals) and 12 (students) out of 28 translationese indicators.

A few features (*pasttense* and *advmod, addit* in student translations) can be described as contributing to **SL/TL-independent translationese**. This group represents the trends in translational behaviour that can be explained by factors other than SL/TL contrast.

For some features, sources and TL reference compare differently in professional and student translations. It may or may not result in changes in their interpretation in terms of investigated translational trends. For example, both professionals and students generated fewer simple sentences than can be expected in the TL. However, professional translators faced a much more challenging situation: their sources had fewer simple sentences than in the TL to start with. Students, on the contrary, did not have to overcome the pressure of the ST as their sources had more simple sentences than in the TL. Nonetheless, students demonstrated a known translational behaviour to use longer sentences with multiple clauses and converted the surplus of simple sentences into the lack of thereof with regard to the TL norm.

The case of additive discourse markers (*addit*) is another example of how ST properties can impact the outcomes of this analysis. Professionals seem to have carried over the additive markers that were seen in the source texts, while students' behaviour was indeed divergent from both ST and TL: students generated surplus additive markers where they were not seen in sources. Overuse of overt linking parentheticals is a known problem of novice writers and translators, who rely on 'crude force', using external markers of cohesion instead of making efforts towards a better sentence structure.

An anticipated outcome of the analysis is seen for analytical passives (*aux:pass*). Both students and professionals avoid carrying over English passive form, which has a corresponding form in Russian. However, professionals view the legitimate isomorphic Russian form as less acceptable, which leads to the lack of it in translations (as compared to non-translations). Both professionals and students seem to make efforts to replace English analytical passives with either of the other three Russian means to express passive relations as can be seen from the frequencies for overall counts for passive forms (see *passives* in Table 5.6). Interestingly, textbooks on practical translations from English into Russian tend to emphasise passives as a typical source of less fluent translation solutions. Some authors make it appear that there are more passive constructions in English than in Russian (see, for example, Zrazhevskaya and Belyaeva, 1972; Belyaev, Popova and Serova, 2010; Borisova, Grinshtein and Protchenko, 2019). We have seen, however, that it is only true for analytical passives.

The strength of association for most features is higher for professional than for student translations. We attribute this to structural simplicity of the source texts, offered to students, on the one hand, and to the difference in the size of the underlying subcorpora, on the other. As demonstrated in Table 5.5, student translations had fewer instances, the smallest mean document length in tokens and in number of sentences, and at the same time the largest STD for both parameters. Smaller corpus size could have contributed to the sparsity of investigated features.

| subcorpus | docs | tokens per doc | tokens per sent | sent per doc |
|---|---|---|---|---|
| student translations | 334 | 646.88 +/-984.07 | 22.39 +/-4.81 | 29.68 +/-40.68 |
| professional translations | 404 | 950.45 +/-282.42 | 22.95 +/-4.03 | 42.44 +/-13.42 |
| non-translations | 497 | 1053.80 +/-539.18 | 20.48 +/-3.04 | 52.39 +/-26.73 |

Table 5.5: Additional corpus parameters for translationese classifications

Besides, student translations were more diverse in the types of translation solutions attempted for typical translation problems. Their renditions were less predictable than stable choices by professional translators, which resulted in a greater spread of values on translationese indicators.

To complete the overview of UD features usefulness for translationese classification, we offer a description of translationese indicators that were specific for detecting either professional or student translations.

Table 5.6 gives contrastive-comparative profiles for the features that were specific for each translation collection in the translationese classification. The table is sorted by F1-score in the respective classification. Equals sign in *means* column indicates no statistical differences between the shorthanded categories. Again, features that were selected as best predictors of student translations against non-translations and had above-chance association with the label are shown in bold.

| | Professional | | | Student | | |
|---|---|---|---|---|---|---|
| feature | LogR | group | means | LogR | group | means |
| Features specific for professional vs non-translation classification | | | | | | |
| mdd | **0.57** | overuse of SL | ref<tgt<src | 0.39 | overuse of SL | ref<tgt<src |
| case | **0.54** | overuse of SL | ref<tgt<src | **0.54** | overuse of SL | ref<tgt<src |
| adp | 0.36 | **useless** | tgt=ref=src | 0.37 | adaptation | tgt=ref<src |
| cont-TTR | 0.36 | adaptation | src<ref=tgt | 0.37 | underuse of TL | tgt<src<ref |
| passives | 0.36 | normalisation | src<tgt<ref | 0.37 | normalisation | src<tgt<ref |
| Features specific for student vs non-translation classification | | | | | | |

| | | | | | | |
|---|---|---|---|---|---|---|
| **parataxis** | 0.37 | adaptation | src<ref=tgt | **0.67** | normalisation | src<tgt<ref |
| **mhd** | **0.63** | SL/TL-indep | ref=src<tgt | **0.63** | SL/TL-indep | ref=src<tgt |
| **numcls** | **0.71** | anglicisation | ref<src<tgt | **0.52** | russification | src<ref<tgt |
| **xcomp** | **0.68** | shining | ref<src=tgt | **0.52** | shining | ref<src=tgt |
| mpred | **0.53** | overuse of SL | ref<tgt<src | 0.44 | overuse of SL | ref<tgt<src |
| aux | 0.42 | overuse of SL | ref<tgt<src | 0.4 | adaptation | tgt=ref<src |
| nummod | 0.46 | shining | src=tgt<ref | 0.38 | adaptation | src<ref=tgt |
| iobj | **0.58** | russification | src<ref<tgt | 0.37 | normalisation | src<tgt<ref |
| infs | 0.39 | russification | src<ref<tgt | 0.37 | adaptation | src<ref=tgt |
| nn | 0.38 | normalisation | src<tgt<ref | 0.37 | russification | src<ref<tgt |
| caus | 0.36 | russification | src<ref<tgt | 0.37 | normalisation | src<tgt<ref |
| cont-dens | 0.36 | normalisation | src<tgt<ref | 0.37 | russification | src<ref<tgt |
| mquantif | 0.36 | shining | ref<src=tgt | 0.37 | **useless** | tgt=ref=src |
| superl | 0.36 | adaptation | tgt=ref<src | 0.37 | adaptation | tgt=ref<src |

Table 5.6: Properties of five and 14 UD features that were unique for professional and student experiments

Quite surprisingly, RFECV selections of features specific for either professional or student translation classification included features that were useless in that particular setting from the point of view of univariate analysis (see adp and mquantif). This might mean that they were employed as part of a multivariate pattern detected by an algorithm.

Some features were not selected for a particular classification despite being well-associated with the class. For example, number of clauses (*numcls*), mean hierarchical distance (*mhd*) and frequency of clausal complements without subjects (*xcomp*) had F1-score over 0.63 when used in a single-feature classifier for professional translations vs reference, but were not selected by RFECV algorithm in that classification. Most likely these features, ignored by RFECV, were collinear with other features in the data. Number of clauses is negatively correlated with ratio of simple sentences.

*Xcomp* was well-correlated with infinitives and number of modal predicates, because it mostly captured typical cases of transferring English modal verbs into Russian, where in non-translations other ways of expressing subjective modality would be used. Example 5.3.1

has a case of positive transfer of a modal predicate, where a more fluent rendition could be Ошибки нельзя исправить. [It is impossible to correct errors].

(5.3.1) [about death penalty] Mistakes can never be rectified.

Ошибки не могут быть исправлены.

'Mistakes cannot be corrected'.

This typical example also demonstrates the multivariate nature of translationese. Following the English pattern in cases like in Example 5.3.1 drives up the counts for analytical passives and *xcomp*.

Among the 13 UD features that were discarded by RFECV as useless in both classifications (see page 158), only one feature – *frequency of non-core verbal argument* (*obl*) – showed reasonably high correlation with the class (F1-score of 0.64 and 0.59 for professional and student classification, respectively). This can be explained by its collinearity with adpositions because oblique nominal objects are introduced by this class of function words.

For consistency, we added the results for full alphabetised list of features in separate tables for each translation variety in Appendix F.

Drawing conclusions from analysis of UD features, we will focus on properties of student translations as the more relevant variety for the subsequent TQ tasks. We found that the outcomes of translationese classifiers on student and professional translations were affected by the differences in the underlying source text collections. Students seem to have worked with shorter and structurally less challenging documents that were more similar to the reference corpus. We can hypothesise that students also generated a wider range of renditions for typical translation problems which made them more diverse as a category and less predictable.

RFECV revealed 42 strong translationese predictors for student translations. Two thirds of them were also found useful to detect professional translations. Only 17 features out of 42 had a reasonably high association with translation/non-translation class labels. The F1-score in a single-feature LogR classifier was in the range of 0.5 to 0.71.

Nine of these individually well-associated 17 features represented a tendency to carry over ST properties into the target text (overuse of ST patterns and shining-through)[8].

[8]inc. shining-through indicators: *acl:relcl, acl, advcl, ccomp, copula, mark, obj, nsubj, xcomp*

Seven more features reflected the tendency to avoid isomorphic (structurally parallel) solutions in favour of the options more typical for the TL (normalisation), which sometimes makes translations over-emphasise TL properties (russification)[9]. One feature (*mean hierarchical distance*) in this set had frequencies dissimilar to both ST and TL. We excluded sentence length and word length from this analysis because these parameters are imposed by the language system, which undermines their usefulness for the proposed analysis, at least in the part of cross-linguistic comparisons.

Analysis of translationese trends should be taken with a grain of salt. Our setup did not allow to test for other possible constraints that affect decision-making in translation apart from SL-TL contrast (e.g. cognitive load, language prestige and established norms). A dramatic decrease in the ratio of simple sentences and changes in the values of the associated properties (most notably, total number of clauses and number of explicit syntactic subjects) is likely to capture a more universal tendency to sentence lengthening and increased syntactic complexity in translations, which is not specific to the given language pair.

### 5.3.2 Best of collgram features

Despite comparatively poor performance of *collgram* features, an attempt was made to identify components with higher predictive power. RFECV algorithm which determines the optimal number of features (instead of setting N a-priori), could not find any features to exclude to achieve better results for either of the translationese classifications. It means that features included in *collgram* set were equally useful/useless.

The poor performance of collocational features indicates that they are unable to effectively pick theoretically expected collocational differences between translations and originally-authored texts. Indeed, the estimated effect size for (statistically significant) differences between categories on most of these features was small. As explained above, we measured the association between a variable and a category label using F1-score from a single-feature *Logistic regression* classifier. In classification on professional translations there were only four features with this score over 0.52 (*bigram-tscore_std, av_bigram-tscore>0, av_trigram-tscore>0, trigram-tscore_std*), while for student translations only

[9] *amod, attrib, fixed, nmod, nnargs, numcls, parataxis*

one feature surpassed F1=0.46 (*trigram-tscore_ std*[10]).

We noticed that values for these most successful collocational features were lower in translations (professional or student) that in non-translations. In fact, lower values for translations are observed for all statistically significant differences in this feature set, except for OOV ratios. No statistical differences were found for features that capture the ratio of negatively-collocated items.

These observations seem to indicate that translations tend to have lower ratios of recognisable collocations of any type (i.e. measured by either NPMI or t-score) than non-translations. The collocations found in translations are less stable and less varied in terms of association strength. The strongest translation predictor in this feature set is OOV ratio, i.e. the proportion of collocations not attested in the non-translated TL.

### 5.3.3 Best of n-gram features

The observations on better-performing features from n-gram feature set, confirmed that, compared to non-translations, translations had significantly higher proportions of bigrams and trigrams not attested in a large TL model (up to 60 and 61% for trigrams in professional and students vs 56% in the comparable non-translations subset). The ratios of OOV was a relatively strong predictor of the category as was the case with collocations: there were more unseen bigrams and trigrams in translations than in non-translations.

We observed that any features based on unigrams performed worse than on bigrams, followed by trigrams. Similarly, ratios of bigrams/trigrams from the bottom-frequency quartiles had less predictive power in either of our translation detection setups than from top-frequency quartiles.

Note that the subset of non-translations (ref) used in the classification experiments was not intersecting with the resource used to train TL n-gram model. Besides, recall that we replaced all proper names and their sequences with a PROPN tag to reduce the expected impact of foreign names on these counts. The values for these features were quite consistent across all documents in each category and had STD of not more than 6%.

Quite unexpectedly, average n-gram LM sentence perplexity (as calculated with *KenLM*

[10]standard deviation for the association scores in each document

library) was higher for non-translations than for translations, and student translations were closer to originally-authored documents (cf. professionals: 3,677.36, students: 4,310.06, non-translations: 5,452.89). It should be noted that sentence perplexity scores varied a lot from sentence to sentence in a document. In non-translations subcorpus, the average STD for a document was the largest and reached 15105.84, three times larger than the mean. This can indicate a higher lexical variety of original texts and a relatively limited, standardised word choice in professional translations. Student translations, on the one hand, might contain more cases of unexpected lexical strings, including spelling errors. This finding is at odds with results reported by (Bizzoni and Lapshinova-Koltunski, 2021), who compared student and professional translations from English into German. They were surprised to find that professional translations had higher values for LM perplexity than student translations.

The sets of features most associated with the category revealed by single-feature classifiers had only two features in common for classifications based on student and professional translations (trioov, trifreq), with student translations being more challenging.

Summarising the performance of abstract lexical features, we can report that student and professional translations shared three *collgram* features identified by RFE (with N=6) as having the highest predictive power (*av_bigram-tscore>0, %bigram-tscore>0, av_bigram-npmi>0*). In the n-gram feature set all six selected features were shared (*bioov, pplex, trifreq, bifreq, trioov, sdpplex*).

In univariate analysis, there were two collgram and two shared n-gram features reasonable well-associated with the category.

The union of outcomes of feature selection and univariate analysis gives a list of 10 best translationese indicators out of the total of 34 features in *collgram* and *n-gram* feature sets.

1. **av_bigram-npmi>0:** average NMPI score for all positively scored bigrams in the document,
2. **av_bigram-tscore>0:** average t-score for all positively scored bigrams in the document,
3. **%bigram-tscore>0:** ratio of all identified bigrams (based on t-score) to the length of the document (in tokens),

4. **trigram-tscore_std:** standard deviation for the association scores in each document,
5. **bifreq:** ratio of bigrams in the top frequency quartile (based on a large non-translated TL corpus),
6. **bioov:** ratio of bigrams absent in the non-translated TL corpus,
7. **trifreq:** ratio of bigrams in the top frequency quartile (based on a large non-translated TL corpus),
8. **trioov:** ratio of trigrams absent in the TL corpus,
9. **pplex:** LM perplexity score averaged across sentences in the document,
10. **sdpplex:** standard deviation of sentence perplexity scores for the document.

Based on our analysis, these features to some extent capture abstract lexical properties of translations, which set them apart from non-translations.

## 5.4 Discussion

This chapter explored the performance of the proposed feature sets, motivated by previous research in contrastive linguistics and translationese studies, in translation detection task. We used two independent parallel corpora to ensure the reliability of our results.

The performance of ML algorithms on hand-engineered translationese features was compared to the results on document-level vectors from a *tf-idf* model and on vectors from two types of embeddings: cross-lingual sentence embeddings (fine-tuned to predict either semantic similarity and MT quality) and word embedding (large dedicate TL model and the most recent multilingual SOTA in representation learning).

**Outcomes of translationese classifications on various representations** The results of translationese classifications on professional and student translations show that both translation varieties are **easily distinguishable from comparable non-translations** in the TL on all attempted representations.

Structural morphosyntactic classifiers (UD) with linear SVM kernel demonstrated **competitive performance** in comparison with content-dependent classifiers (*tf-idf* and embedding models). The best result on UD features after feature selection for students was 91.06% and for professionals 92.04% (cf. $F1 = 98.36$ for *mdeberta3*). The superior performance of the latter was expected as they had access to possible topical differences

between the document categories. Although additional efforts were made to ensure that non-translated TL subcorpus was comparable to translations in terms of functional register (see Section 3.2.3), translations and non-translation might have retained considerable topical content variation. SVM translationese classification results were replicated by a neural classifier.

The abstract lexical features arranged into *collgram* and *n-gram* feature sets **were less successful** in translation detection tasks than morphosyntactic features. The poor performance of collocational features indicates that they are unable to effectively pick theoretically expected collocational differences between translations and originally-authored texts. The effect size, measured as the association of each continuous variable and binary categorical labels using *Logistic regression F1-score*, was small on most of these features for differences between categories. There were only four features (out of the total of 34) with the effect size score over 0.52 in translationese classification on professional translation, for student translations only one feature surpassed $F1 = 0.46$. Although the performance on *collgram* and *n-gram* feature was well above the chance level, they did not provide statistically significant gains when combined with more successful morphosyntactic features.

Translationese classifications on professional and student translations yielded mildly surprising results when compared to each other. Against expectations, professionals were a bit more distinct from non-translations than students. Admittedly, the only statistically significant result between translationese classifications on student and professional translations was seen for *stsb-xlm-r-m* for both SVM and neural classifier. We explained this oddity by the scale of semantic and structural differences between Russian non-translations and English sources for student and professional translations, using PCA visualisations and feature analysis. Besides, student translational solutions were less stable and patterned. They ranged from carrying over English structures where possible to (sometimes unnecessary and unsuccessful) re-working of the entire sentences in an attempt to avoid following the ST. However, the lack of writing skills in Russian could lead to paradoxically poor and perplexing results in the latter case. A student translation in Example 5.4.1 ignores a typical Russian way to express the ST idea (notice the introduction of Russian passive) and uses an English-sounding modal predicate.

(5.4.1) ...270 million have no access to health services.

...270 миллионов *не могут* воспользоваться медицинскими услугами.

'270 million cannot make use of medical services'.

Possible variant: 270 миллионов лишены возможности получать медицинскую помощь. [270 million (people) are deprived of ability to get medical help.]

Example 5.4.2 presents a typical case of excessively wordy translation, which unnecessarily introduces a string of two de-verbal nouns, which is a known indicator of low writing skills tackled in translation textbooks. At the same time this translation ignores differences between English and Russian in organising the information flow from theme to rheme and fails to move the emphasised non-core verbal argument to the sentence end.

(5.4.2) Even among those with dementia, effective treatment [of depression] often reduces symptoms.

Даже среди людей страдающих старческим слабоумием эффективное лечение способно приводить к *уменьшению выраженности* симптомов.

'Even among people suffering from dementia effective treatment is able to lead to *a decrease in manifestation* of symptoms'.

Possible variant: Известно, что эффективное лечение часто сокращает симптомы [депрессии] даже у пациентов, страдающих слаюлумием. [It is known that effective treatment often reduces symptoms (of depression) even in patients with dementia.]

Tracing the differences in between student and professional translations emphasises that **translationese distinctions are contingent on the properties of ST rather than SL**. Translations based on slightly different ST collections can return a bit different translationese-related results against the same non-translations sample. It follows that variation in ST is a confounding factor in translationese research, and ideally, it should be controlled (or accounted) for in an experimental setup (for example, professional and student translations should be obtained for the same STs, which was not possible in our research). In our setting we found that students translated documents that were more similar to the reference corpus than sources in the professional subcorpus. As a result, students were a bit more difficult to distinguish from the reference corpus.

We used PCA analysis and visualisation of its results on two translation collections to

shed light on the **dissimilar aspects of documents relatedness that were captured by various representations** employed in this thesis. Based on the type of representations and the outcomes of the comparison between two parallel corpora against non-translations, we can tentatively conclude that UD features, cross-lingual *TQmono-m* model and monolingual Russian model (*ruRoberta*) captured targeted translationese distinctions, while multilingual *mdeberta3* and cross-lingual sentence-embedding XLM-R model (*stsb-xlm-r-m*), achieved the same results in translation detection tasks looking at topical differences between the documents.

Although neural features from embedding models are said to capture complex non-local syntactic and semantic information that discrete hand-engineered features can hardly encode, it is difficult to be sure that they compare documents along the targeted aspects (translation vs non-translation properties in this case).

**Description of translationese indicators** Section 5.3 was focused on feature analysis. We aimed at the identification of features and their subsets that captured translationese best. If our hypothesis stands, these features would emerge as useful quality predictors. To this end, we used feature selection as a method of **multivariate analysis**. It was complemented with **univariate analysis** of individual features using single-feature classifiers and statistical tests. We identified the best translationese indicators as features selected by RFECV algorithm with the highest performance on the task in a single-feature classifier.

Testing our features on two parallel corpora helped establishing their robustness in detecting translations. As mentioned above, structural UD features were the more useful feature set for this task. Reducing the UD feature space by RFECV led to a small improvement in both classifications, with student translations relying on a larger subset to ensure maximum gain in performance. As expected, many of these translationese indicators were shared between the translation varieties and included 47% (28 features) of the original feature set of 60.

There were thirteen features that were not selected by RFECV for either of the classifications. Additional statistical tests demonstrated that an overwhelming majority of them were sparse and weakly-associated with the labels.

This chapter also presented a **method to describe translationese predictors as contributing to one of the three translational behaviours** that can be explained

by the influence of either ST or the TL norm. On the better-performing UD feature set for student translations, also limited to the features that had a reasonable association with the class (17 features), this approach allowed to establish that student translations deviated from the TL norm along two major directions. Nine features were reproducing ST patterns and seven features reflected active attempts to reconcile the differences between SL and TL in favour of the latter. There was only one feature that had values deviating from both SL and TL.

Analysis of the features that were brought up as translationese indicators confirmed some common anticipations from translation theory and practical textbooks on English-to-Russian translation. In particular, *translations had higher frequencies of additive discourse markers, analytical passives, copula verbs, modal predicates, personal pronouns, finite verbs and determiners* than non-translations. The most prominent property of translations against non-translations was longer and more complex sentences.

We see two major limitations of this analysis. First, it does not account for other known factors that might result in translationese deviations, except for SL-TL contrast. These factors might include other known tendencies in translation, most importantly explicitation, and lack of developed writing skills in the TL. The results of a detailed manual analysis based on a very similar subset of student translations from the RusLTC, reported in (Kunilovskaya et al., 2022) as well as examples above indicate that the properties of student translations can be shaped by 'unforced' errors, when students fail to generate a fluent output due to limited TL competence. Second, although the analysis is based on document- and sentence-level aligned corpora, we did not link the individual items within sentences. It is likely that specific frequencies in translations were not generated directly by the frequencies of comparable items in the ST.

The translationese classification performance on abstract lexical features did not live up to our expectation, inspired by the literature. Nonetheless, we were able to see that the best translationese predictors in both feature sets were ratios of OOV, which indicates the presence of 'strange strings' (that were not proper names!). Bigrams seemed to work better than trigrams, and differences in the ratios of items from the top frequency quartile were more useful than from the bottom quartile. We were unable to establish any differences in the performance of features based on NPMI and t-score. However, generally, our experiments confirmed that translations had a smaller variety of collgrams

in terms of association scores, and the average association scores for identified collgrams for translations was lower than for non-translations. The most surprising finding was that a n-gram LM trained on a large corpus of Russian newspaper texts returned the highest perplexity score for a (non-intersecting) subset of non-translations than for translations. It is likely that translations were more patterned and repetitive in expression, despite the presence of 'strange strings'. However, the comparison of the two translation varieties was in line with what was expected: students were more perplexing for the model than professionals.

# Chapter 6 Experiments on Varieties and Quality

This chapter reports the results of experiments using linguistically-motivated translationese-aware features to distinguish professional varieties of translations and to estimate human translation quality on various types of human quality assessment. Importantly, these results are compared to the performance of several other representations to indicate the learnability of the available ontological labels and human judgment.

The results are reported in two groups of experiments run for the same types of document representations.

First, Sections 6.1.1 and 6.1.2 have the results of binary classifications for student vs professional translations, and for documents with *bad* and *good* quality labels, which were obtained as described in Section 4.4.1. Second, in Section 6.2, we explore the learnability of document-level and sentence-level continuous quality scores generated (i) from error annotation (as discussed in Section 4.4.2) and (ii) from direct assessment (see Section 4.4.3). The experiments on quality scores, formulated as regression tasks are extended to sentence level, because both types of quality scores (unlike labels) were originally obtained for sentences and then averaged to get document scores. On the one hand, sentence-level tasks are expected to be more challenging for most of manual frequency-based features, but on the other hand they should be more natural for embeddings and less limiting in terms of the size of the training data. Additionally, scores from two different manual assessment methods (error annotation vs perceived quality) might offer dissimilar perspectives of translation quality. Last, but not least, scores from error annotation allow to compare applicability of proposed representations for major theoretical quality aspects – accuracy and fluency.

The classification tasks in Section 6.1 rely on the same experimental setup as described in Section 5.1 with regard to learning methods, most alternative representation methods and feature analysis. However, we open this section with details on the datasets and on additional representation for varieties and quality classifications.

The specifics of regression experiments appear in the second part of the chapter, before we present the respective results (see Section 6.2.1).

## 6.1 Classification Tasks

There are a few preliminary notes on the classification experiments reported in this section.

**Datasets** Classification experiments were run on the same dataset as translationese classifications. However, it does not mean that student translations were represented only by translations with either good or bad label. In fact, 334 student translations and 360 quality-labelled translations had only 25 instances in intersection (six labelled as *good* and 19 as *bad*).

The binary labels used the main experiments reported in this chapter are:

**stu vs pro:** categories in a classification to capture professionalism;

**bad vs good:** categorial document-level quality labels from holistic assessment of student translations verified in a separate annotation effort as described in Section 4.4.1.

**QuEst++ baseline** In addition to manual features investigated in this thesis, we implemented 61 of 78 available document-level QuEst++ features (*quest61*) to ensure comparability of our results with previous research. We had to exclude QuEst++ features which reflected the size of the documents – number of tokens in the source document (#1001), number of tokens in the target document (#1002) and number of sentences in the target document (#9801) because they would be a simple give-away for student and professional translations in our collections (see additional parameters of the respective document categories in Table 5.5 on page 165).

### 6.1.1 Translation Varieties

In this section we look at the differences between professional translation varieties, when classified against each other. This opposition is interpreted as a proxy for quality on the assumption that student translations are inferior to professional due to the lack of experience (see Section 3.1.2). A useful reminder is that for this experiment the student translations were selected randomly from the available sets of multiple translations for each source text and they might come from various quality grades.

In the previous chapter we have seen that students were not easier to tear apart from non-translations than professionals. On the contrary, in terms of returned F1-scores, the

results for translationese classifications on professional translations were slightly higher than on student translations for representations that capture translation-related aspects of language, i.e. on all manual feature sets and on sentence embeddings that were fine-tuned on translation quality (*TQmono-m*). However, on *tf-idf* and on vectors from generic word-embedding models (*ruRoberta* and *mdeberta3*) the results for student translations were higher.

**Experimental results** Table 6.1 has the results on classification of translation varieties. We explicitly added results on the content-dependent classifiers that were not shown before, but described in *Alternative document representations* in Section 5.1 on page 142 (*tfidf, stsb-xlm-r-m, TQmono-m, ruRoberta*).

| rep | Accuracy | Precision | Recall | F1 |
|---|---|---|---|---|
| chance | 49.59 (+/-3.47) | 49.45 (+/-3.48) | 49.45 (+/-3.51) | 49.39 (+/-3.49) |
| UD | 76.55 (+/-5.15) | 77.05 (+/-5.39) | 76.32 (+/-4.77) | 76.24 (+/-5.01) |
| collgram | 65.02 (+/-5.29) | 65.50 (+/-6.24) | 63.70 (+/-5.27) | 63.25 (+/-5.74) |
| ngram | 66.80 (+/-5.64) | 67.11 (+/-6.12) | 66.32 (+/-5.60) | 66.07 (+/-5.74) |
| all | 79.67 (+/-3.50) | 79.85 (+/-3.48) | 79.46 (+/-3.32) | 79.43 (+/-3.46) |
| quest61 | 83.20 (+/-2.85) | 83.20 (+/-2.91) | 82.96 (+/-2.85) | **83.00** (+/-2.87) |
| tfidf | 89.69 (+/-2.70) | 89.69 (+/-2.71) | 89.60 (+/-2.76) | 89.59 (+/-2.72) |
| stsb-xlm-r-m | 74.12 (+/-3.86) | 74.55 (+/-4.05) | 73.49 (+/-3.87) | 73.49 (+/-3.98) |
| TQmono-m | 84.28 (+/-3.70) | 84.87 (+/-3.45) | 83.68 (+/-3.98) | 83.89 (+/-3.89) |
| ruRoberta | 87.26 (+/-4.22) | 87.73 (+/-4.40) | 86.95 (+/-4.25) | 87.05 (+/-4.29) |
| mdeberta3 | 89.03 (+/-2.87) | 89.40 (+/-2.85) | 88.69 (+/-2.95) | 88.85 (+/-2.93) |

Table 6.1: Classification result for translation professionalism labels (student vs professional)

A glance at Table 6.1 suggests that the best performing representation on this task among manual features was the full QuEst++ feature set ($F1 = 83.28$). Among content-dependent classifiers *tf-idf* was not significantly outperformed by any other representation ($F1 = 89.59$). If we consider only statistically significant differences between representations for both SVM and neural classifier, we can observe the following.

The relatively poor performance of translationese-aware UD feature sets against MT-quality-aware QuEst++ feature set (differences are statistically significant) suggest that the translationese-related distinctions between translation varieties were weaker than

those captured by MT quality indicators (see feature analysis below).

Further on, *tf-idf* was statistically superior to QuEst++. This means that the topical differences between the parallel corpora were even stronger than quality-related or translationese distinctions.

This line of interpretation is supported by the similarities in performance between QuEst++ and *TQmono-m*, the two approaches designed to capture quality in MT, and between *tf-idf* and *mdeberta3*, which are thought to be focused on token-based properties (semantics) of documents.

The outcomes of significance testing between the representations is summarised in Table 6.2. A hyphen indicates no differences, an asterisk stands for statistically significant differences as measured by t-test at the confidence level of 5%.

| | UD | collgr | ngram | all | quest61 | tfidf | stsb-xlm | TQmono | ruRoberta |
|---|---|---|---|---|---|---|---|---|---|
| UD | | | | | | | | | |
| collgram | * * | | | | | | | | |
| ngram | * * | - - | | | | | | | |
| all | - - | * * | * * | | | | | | |
| quest61 | * * | * * | * * | * - | | | | | |
| tfidf | * - | * * | * * | * - | * * | | | | |
| stsb-xlm | - - | * * | * * | * - | * * | * - | | | |
| TQmono | * * | * * | * * | * - | - - | * - | * * | | |
| ruRoberta | * * | * * | * * | * * | * * | * * | * * | - * | |
| mdeberta3 | * * | * * | * * | * * | * * | * * | * * | * * | - - |

Table 6.2: Professionalism: Results on pair-wise significance testing on SVM (left marker) and neural classifier (right marker)

To further explore the success of QuEst++ features we ran RFECV and analysed the 25 selected features. This selection improved the performance of the classifier by 3% against the entire feature set in terms of F1-score (86.17% vs 83.00%, the difference was statistically significant).

We noticed that 15 of 25 features selected by the algorithm describe the SL side (complexity features), seven features reflect target properties (fluency features), and 3 features are ratios of the two (adequacy features). The five most prominent ST-related QuEst++ features include *percentage of punctuation marks in the source* (#1074), *source sentence perplexity without end of sentence marker* (#1011), lemma repetition in

the source document (#9991), average number of translations per source word in the document (#1024) and average source token length (#1006). Given that the sources of student and professional translations are different, it is not surprising that this feature set works well. However, it is difficult to interpret the distinctions as related to professionalism or quality. For this feature set, which explicitly uses ST complexity features, either input to the translation process or the method of translation (type of human subjects, in our case) need to be the same.

The prominent fluency and adequacy features that can be viewed as relatively independent of the properties of ST and can be interpreted as related to quality in our experimental setup. Selected *fluency features* were *number of occurrences of the target word within the target hypothesis or TTR* (# 1015), *log probability of the target* (# 1012), *perplexity of the target document without end of sentence markers* (# 1014). The log probabilities of student translations were on average lower than for professional translations and perplexity scores were higher for students. We can hypothesize that student translations were more surprising than professionals for the Russian LM trained on a comparable corpus. This confirms our findings in Section 5.3.3 (page 169) and can be indicative of the presence of 'strange strings' and awkward word choice, a common problem in student translations, accounting for up to 30% of all annotated errors (see Figure 4.4, page 128). As expected TTR was lower for student translations. It is a well-known translationese indicator: translations tend to have lower TTR than non-translations.

The *adequacy features* selected by RFECV included only three features: *ratio of number of tokens in source and target* (#1003), *ratio of number of tokens in target and source* (#1004) and *ratio noun repetition between target and source documents* (#9996). The comparison of average values for professional and student translations showed that professional translations were more wordy but less repetitive.

Despite QuEst++ feature set was not entirely fit for our purposes, the relevant parts of it confirmed that **student translations were indeed different from professional translations in terms of quality-related features of fluency and adequacy**.

Turning to the analysis of best performing UD feature, we can report that a classifier limited to 15 RFECV-selected features improved the performance on this task by three percentage points from F1-score 76.24% to 79.39%. Only six of these strong predictors of professional varieties in translations were also among strong translationese indicators.

This list included frequencies of relative clauses, clausal complements, copula verbs, negative particles, function words indicating the head of a subordinate clause, subordinating conjunctions.

Other six features come from the subsets of translationese indicators that were specific only to professional (prepositions, function of analytical case markers (mostly prepositions), type-to-token ratio) or student translations (asyndetically introduced elements, indirect objects, nouns).

The third subset (appositional modifiers of nouns, coordinating conjunctions, syntactic conjuncts) included features that were not selected in either of the translationese classifications.

Based on the performance of single-feature classifier (if there were statistical distinctions between the categories in the first place), these features cannot be viewed as capturing large distributional dissimilarities between student and professional translations. Only seven out of 15 selected features had statistically different values between the categories. The highest effect size among them was observed for *parataxis*: LogReg F1-score for this feature was estimated at a very modest level of 34%.

We have already seen in the previous chapter that both translation varieties deviate from the expected TL norm along the same dimensions suggestive of similar patterns in translational behaviour.

Despite professionalism was captured fairly well by UD features (F1 = 79.39% on the best-performing subset), the differences between the varieties can hardly be explained by translationese. If anything, some expected translationese trends were better expressed in professional translations (based on the features with statistical differences): they had more markers of dependent clauses, clausal complements, epistemic markers, modal predicates as well as fewer nouns. In English-to-Russian translation these feature might indicate carrying over the English patterns rather than adopting to the TL norm. Some strong translationese predictors (*acl:relcl, copula, determiners, finite verbs*) did not have statistical differences between the categories. On the other hand, professional translations had more varied vocabulary evidenced by higher TTR, more indirect objects and negative particles, fewer analytical passives. These properties are indicative of stronger normalisation trends along these dimensions.

| | documents | sentences | tokens | | tokens per doc | |
|---|---|---|---|---|---|---|
| | | | EN | RU | EN | RU |
| bad | 183 | 4 144 | 93,844 | 87,894 | 512.81 +/-117.39 | 480.30 +/-114.76 |
| good | 177 | 4 070 | 89,536 | 85,560 | 505.85 +/-108.44 | 483.39 +/-111.83 |

Table 6.3: Basic parameters of parallel subcorpus with binary quality labels (after preprocessing and lemmatisation)

### 6.1.2 Binary Quality

This section presents the results of the same analytical procedure applied to document-level quality labels. Our major question is whether the identified translationese indicators consistently contribute to predicting 'good' and 'bad' translation.

The dataset with binary quality labels is summarised in Table 6.3. It can be seen that we have a reasonably balanced corpus, with a bit fewer good translations. Recall from Section 4.4 that these documents come from sets of multiple student translations for 57 English mass-media source texts, and include from one to four translations in each quality category for each ST.

Table 6.4 reports SVM outcomes for 10 representation of student translations annotated with binary quality labels. The results for the neural classifier appear in Appendix E. Out of 10 representations attempted for this classification, statistically significant differences between SVM and neural classifier were seen only on *stsb-xlm-r-m* vectors, where SVM outperformed the neural model. As before, below we compare the performance of various representations based on the agreed outcomes of SVM and neural model.

| rep | Accuracy | Precision | Recall | F1 |
|---|---|---|---|---|
| chance | 51.67 (+/-7.78) | 51.67 (+/-7.80) | 51.67 (+/-7.78) | 51.63 (+/-7.78) |
| UD | 61.39 (+/-5.75) | 61.60 (+/-5.81) | 61.29 (+/-5.82) | 61.00 (+/-5.93) |
| collgram | 57.78 (+/-9.53) | 58.04 (+/-9.99) | 57.87 (+/-9.73) | 57.36 (+/-9.81) |
| ngram | 47.22 (+/-4.48) | 47.29 (+/-4.53) | 47.42 (+/-4.41) | 46.73 (+/-4.51) |
| **all** | 62.50 (+/-7.68) | 62.70 (+/-7.92) | 62.45 (+/-7.76) | **61.99** (+/-8.21) |
| quest61 | 47.22 (+/-6.92) | 47.23 (+/-6.96) | 47.28 (+/-6.91) | 46.96 (+/-6.92) |
| tfidf | 59.17 (+/-6.34) | 59.19 (+/-6.51) | 59.11 (+/-6.47) | 58.92 (+/-6.58) |
| stsb-xlm-r-m | 61.67 (+/-9.92) | 61.70 (+/-9.94) | 61.66 (+/-9.93) | 61.61 (+/-9.93) |
| TQmono-m | 63.06 (+/-6.22) | 63.42 (+/-6.10) | 63.11 (+/-6.01) | 62.77 (+/-6.34) |
| ruRoberta | 75.00 (+/-7.14) | 75.50 (+/-7.42) | 75.02 (+/-7.14) | 74.89 (+/-7.13) |
| **mdeberta3** | 78.33 (+/-6.89) | 78.88 (+/-6.48) | 78.34 (+/-6.93) | 78.14 (+/-7.13) |

Table 6.4: Binary quality classification results (neural classifier)

In this classification, UD perform significantly better than all other hand-engineered features, including features from QuEst++ framework. The concatenation of UD with collgram- and unigram-based features did not yield a significant increase in performance.

This task proved much more challenging for content-dependent classifiers than translation detection task or classification of professional varieties. It is not surprising as the classifiers were faced with a content-wise much more homogeneous set of documents, in fact, with multiple translations of the same source text.

There were no statistically significant differences between the performance of UD features and *tf-idf* or vectors from two sentence embedding models (*stsb-xlm-r-m* and *TQmono-m*). However, representations produced as triple aggregation of word vectors using *ruRoberta* and *mdeberta3* models, remained unbeaten in absolute terms, with no difference between them, as was the case with all other classifications. The results of statistical significance tests between representations are presented in Table 6.5. In each pair of markers in the table, the first one is for SVM and the second – for the neural classifier; asterisk indicates statistical significance of observed differences.

Hand-crafted UD features designed to capture translationese (and successful in that task as we have seen in Chapter 5) were secondary only to the recent large pre-trained word embeddings in the task of document-level human translation quality estimation. They achieved $F1 = 61.00$ and $F1 = 78.14$, respectively. Admittedly, their performance was very limited in the proposed binary classification setup, but it was still above the

random baseline, which was not the case for alternative representations.

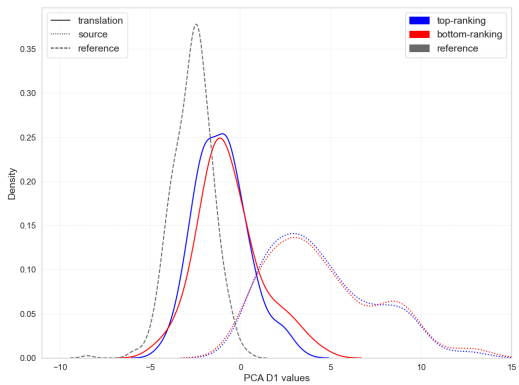
| | UD | collgr | ngr | all | quest61 | tfidf | stsb-xlm | TQm | ruR |
|---|---|---|---|---|---|---|---|---|---|
| UD | | | | | | | | | |
| collgram | - - | | | | | | | | |
| ngram | * - | * - | | | | | | | |
| all | - - | - - | * - | | | | | | |
| quest61 | * * | * - | - - | * * | | | | | |
| tfidf | - - | - - | * - | - - | * * | | | | |
| stsb-xlm | - - | - - | * - | - - | * - | - - | | | |
| TQmono | - - | - * | * * | - - | * * | - * | - * | | |
| ruRoberta | * * | * * | * * | * * | * * | * * | * * | * * | |
| mdeberta3 | * * | * * | * * | * * | * * | * * | * * | * * | - - |

Table 6.5: Quality labels: Results on pair-wise significance testing on SVM (left marker) and neural classifier (right marker)

Good performance of embedding models is reassuring, because it means that the quality distinctions annotated in our dataset were actually related to objective properties of the texts, and not as subjective as it was commonly believed with regard of human annotation of quality at document level.

Although we cannot be entirely sure which linguistic properties of translations were captured by the best performing *mdeberta3* embedding, lack of topical variation between good and bad translations in this experimental setup, allows a cautious conclusion that the vectors actually reflect quality-related aspects of language.

UD features PCA D1




*mdeberta3* PCA D1

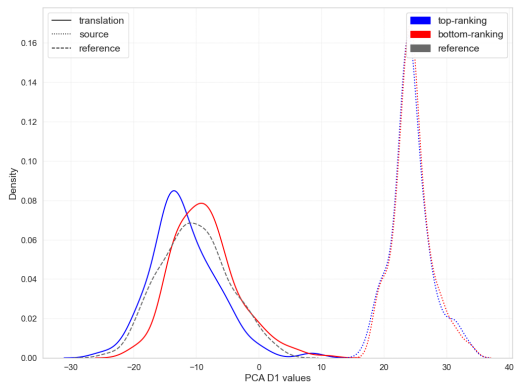


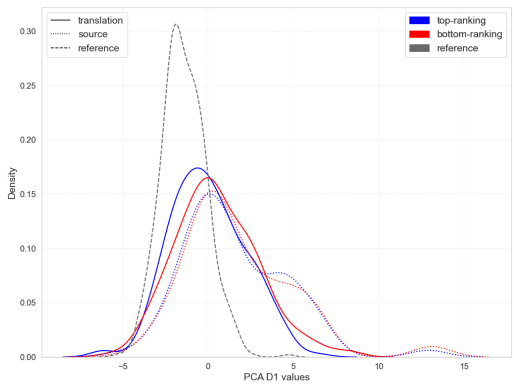

Figure 6.1: Variability of PCA D1 values in quality-annotated documents (along with their sources and non-translations)

The panels in Figure 6.1 reflect the differences between quality categories as captured by the full UD feature set and *mdeberta3* model. UD features clearly put student translations in the gap between source and target languages, with top-ranking translations being a bit shifted towards the non-translated TL side of the continuum. Vectors from *mdeberta3*, a multilingual model, primarily capture the language distinction at D1. This squeezes the Russian language varieties into the same location on the plot and overshadows the differences between them. Nonetheless, top-ranking translations are clearly further away from sources, than low-ranking translations. This might mean that top-ranking translations are more distant from their sources and more adopted to the TL norm.

In turn, this suggests that the quality distinctions captured by *mdeberta3* are, in fact, related to translationese, interpreted as the amount of deviations from the TL norm, especially on the axis capturing language contrast.

However, as we have seen in the feature analysis for student translations in Section 5.3.1, shining-through does not exhaust translationese manifestations. Plotting the PCA results for just the best performing subset of UD features eliminates the distinctions between languages. This might mean that many of the 26 selected features (see below) are not necessarily shining-through indicators. Translation quality is partly captured along other dimensions of translated language, not the one defined by the contrast between languages.

Top 26 UD features PCA D1




*mdeberta3* PCA D1

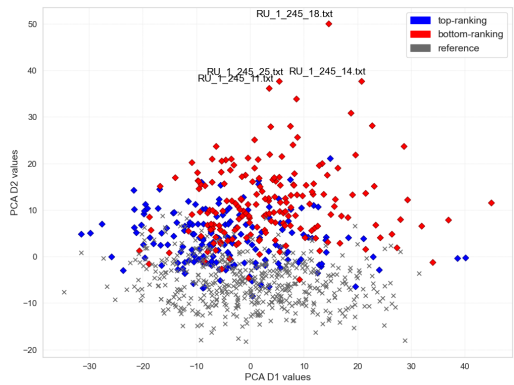



Figure 6.2: Variability of PCA D1 values in quality-annotated documents (no source texts)

The right-hand panel in Figure 6.2 demonstrates that quality parameter of the dataset is captured on the orthogonal PCA dimension (D2): the lower-ranking translations (in

red) can be identified as more distant from TL norm (grey cloud).

**Feature analysis** Almost half of 26 most contributing UD features selected by RFECV in the quality learning task were in the intersection with strong translationese indicators identified in Section 5.3.1. This selection significantly improved the performance of the SVM classifier: F1-score went up by 8% from 61.0 to 68.9%.

Table 6.6 aggregates parameters that characterise each feature in this selection from the point of view of: (1) ability to predict quality categories alone (for features with statistically significant differences between the categories, otherwise '–' in *LogR* column), (2) comparison of mean feature values across the categories, (3) SVM weights assigned to each feature and (4) translational trend in 'bad' translations. In the latter, we showed the hypothesised trend in bold if there were significant differences between 'good' and 'bad' categories. The regular font is used if there were significant differences between some of the three document categories involved in the comparison (src, bad, ref). In cases where there were no distinctions between the three categories, it was difficult to talk about any trend (–). The table is sorted by SVM weight. Features that were also identified as strong translationese indicators in previous analysis (see Section 5.3.1) are shown in bold in *feature* column. Translationese features that were specific for either professional or student translations are shown in green and orange, respectively, as they were colour-coded in Section 5.2. There were only three quality indicators that were not associated with translationese: *frequency of syntactic conjuncts (conj), appositive constructions (appos) and possessive pronouns (possp)*. In *means* column we used <= and => signs to signal the comparison between categories, even in the absence of statistically significant differences.

| feature | LogR | means | SVM weight | trend in 'bad' |
|---|---|---|---|---|
| **sentlength** | – | bad=>good | 1.616 | NA |
| **nsubj** | 0.13 +/-0.12 | bad>src>good>ref | 1.06 | **anglicisation** |
| **amod** | – | bad=>good=>ref>src | 0.857 | adaptation |
| numcls | – | bad>good>src=>ref | 0.713 | SL/TL indep |
| **attrib** | – | bad=>good=>ref>src | 0.623 | adaptation |
| **acl:relcl** | 0.13 +/-0.09 | src>bad>good>ref | 0.615 | **shining** |
| **simple** | – | src>ref=>bad=>good* | 0.563 | adaptation |

| | | | | |
|---|---|---|---|---|
| **advcl** | 0.13 +/-0.07 | src>bad>good>ref | 0.521 | **shining** |
| **aux:pass** | – | src>bad=>ref=>good | 0.491 | adaptation |
| **deverbals** | – | bad=>good=>ref>src | 0.477 | adaptation |
| **obj** | – | src>bad=>good>ref | 0.44 | shining |
| content_TTR | 0.00 +/-0.00 | bad<good<=src<=ref | 0.416 | **SL/TL indep** |
| **mark** | – | src>bad=>good>ref | 0.409 | shining |
| parataxis | – | src<bad<=good<ref | 0.397 | underuse of TL |
| conj | – | bad=>good>src=>ref | 0.389 | SL/TL indep |
| passives | 0.05 +/-0.08 | src<bad<good<ref | 0.32 | **underuse of TL** |
| iobj | 0.05 +/-0.13 | src<bad<good<ref | 0.319 | **underuse of TL** |
| nummod | – | ref=>bad=>good>=src* | 0.318 | – |
| mdd | – | ref<bad<=good<src* | 0.316 | normalisation |
| **pverbals** | – | ref>bad=>good>src* | 0.289 | normalisation |
| adp | – | src=>ref=>bad=>good | 0.254 | – |
| mquantif | – | bad<=good<=ref<src | 0.244 | underuse of TL |
| appos | – | bad=>good=>ref=>src | 0.23 | – |
| possp | – | src<ref<=bad<=good* | 0.202 | adaptation |
| caus | – | ref<=src<=bad<=good* | 0.195 | – |
| mpred | – | src>bad=>good>ref | 0.116 | shining |
| copula | 0.14 +/-0.12 | src>bad>good>ref | | **shining** |
| finites | 0.09 +/-0.13 | src>bad>good>ref | | **shining** |
| infs | 0.08 +/-0.09 | src<bad<good<ref | | **shining** |

Table 6.6: Features selected by RFECV with results of significance testing and comparison between quality categories

The last three features in Table 6.6 had significant differences between the quality categories in univariate analysis, but they were not selected as useful in quality classification (*copula verbs, verbs in finite form and infinitives*). Copula verbs and finite verb forms were among shared translationese indicators, while frequency of infinitives was useful to separate student translations from non-translations.

As can be seen from Table 6.6, SVM classifier operated in the feature space where

only a few features had statistically significant distinctions. All features had very modest association with the labels, too. It is not surprising as this experiment was run on a relatively small dataset. Nonetheless, it can be seen that in cases of significant difference between quality categories, *bad* translations mostly passively follow SL patterns. This behaviour consists in either carrying over the items with higher SL frequencies (anglicisation, shining-through) or in lack of ability to come up with preferred TL patterns, where the frequencies in TL are higher than in the SL (*underuse of TL*). Most comparisons (even if insignificant) showed that *good* translations were almost always closer to the TL reference than *bad* translations (the six exceptions from this observations are marked with an asterisk).

Figure 6.3 visualises the contribution of each feature to the improved classifier.

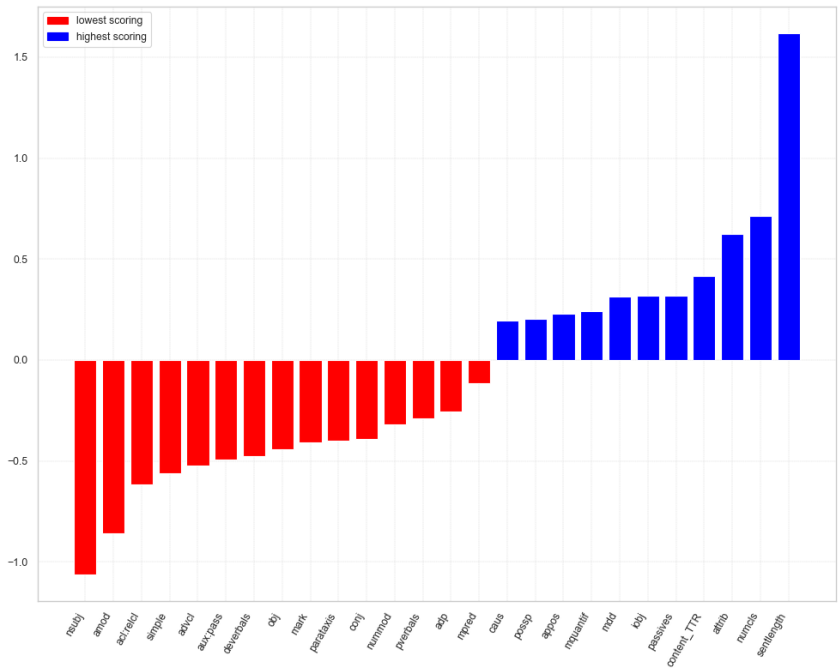



Figure 6.3: Weights loading the decision towards each category for 26 RFECV-selected items

We omitted the analysis of QuEst++ features on this dataset due to its poor performance, which was even lower than the random baseline.

Concluding the observations from the experiments on document labels, let us see whether there were commonalities between the outcomes of professionalism and quality classifications.

Despite the classification results on UD features were higher in professional variety classification than in translation quality classification (79.39% vs 68.9% on the best-performing subsets of features), we would argue that translationese is more related to quality than to professionalism.

In Sections 6.1.1 (page 182), we have shown that some translationese features were actually more typical to professional translations than to student translations. This finding was against our expectations, but helped the classifier. On the contrary, quite in line with the expected, the properties of *bad* translations on most translationese features indicated that these translations had major deviations from the expected TL norm.

It means that distinctions between professional varieties and quality categories run along different directions. **Professional translations do not necessarily exhibit less translationese than student translations, while lower-ranking translations do demonstrate stronger translationese tendencies.** In particular, *bad* translations feature longer sentences, more complex sentence structure (more clauses, especially relative clauses), lower TTR, more analytical passives and fewer passive constructions, more nouns functioning as subjects, more modal predicates, more verbal forms, including overuse of copula verbs, de-verbal nouns and participles.

## 6.2 Regression Tasks

The purpose of this section is to compare the performance of SVM on document-level datasets with two types of quality scores: a score derived from error annotation and a score obtained through direct assessment.

We want to find out which approach to benchmarking quality by human judges is better aligned with the objective properties of translations as captured by our representations, with a focus on the established translationese indicators.

### 6.2.1 Experimental Setup

**Datasets** The input for the quality prediction tasks, formulated as a regression problem was stored in a separate table of the structure similar to the dataset with document-level labels. It included instances in rows and features in columns. All tested scores, sentence-aligned documents and their lemmatised versions were added as columns as

well. The intersection between quality-labelled and quality-scored student translations included about 17% of the total number of unique student translations used in this thesis (206 of 1222 documents). Therefore, the datasets were kept separate.

In brief, the error-annotated subcorpus consists of 553 parallel documents, 140 of which are also annotated with DA scores. Both collections include multiple translations for 46 and 30 STs, respectively. The quality scores were generated as described on page 131 in Section 4.4.2 for error-based scores and on page 137 in Section 4.4.3 for scores from direct assessment. In particular:

**accuracy:** frequency of content errors weighted for error categories and normalised to the ST word count;

**fluency:** frequency of language errors weighted for error categories and normalised to the ST word count;

**tq:** overall quality score and its variant (**scaled_tq**) transformed into the 0-100 range;

**da_mean:** document-level scores from DA averaged from sentence-level and their z-transformed variant (**da_zmean**)

**Representations** We used the same representations as for the labelled datasets. Given the poor performance on *collgram*- and *ngram*-based features and on *stsb-xlm-r-m* vectors in the classification tasks, we omit them from the results in regression experiments reported in this section.

**Learning algorithm** For translated documents with quality scores, the ML task is formulated as a regression problem. We learn the target scores using Support Vector Regressor (SVR) as implemented in *scikit-learn*, with the library default settings. In particular, the algorithm uses *radial basis function (rbf)* kernel, with *gamma* parameter set to 'scale', $C$ to 1.0 and *epsilon* to 0.1. We omit the results for the neural regressor, because they are not different from SVR performance in a principled way as was the case in Sections 5.2, 6.1.1 and 6.1.2.

The performance of SVR is reported in terms of Spearman's rank correlation coefficient (r) and RMSE. The former is selected because the quality scores are not normally distributed especially at sentence level (see Figures 4.7 and 4.8 at page 132 in Chapter 4). RMSE

helps to keep track of the magnitude of errors in predicted scores. This metric is typically used in the quality estimation shared tasks by WMT.

**Feature analysis methods** The strength of the relationship between each variable (feature) and target labels was measured using 5-fold cross-validated *Linear regression* which returned Spearman's correlation coefficient for continuous response variable (scores) in a regression task.

### 6.2.2 Document Level: Error-based Scores

Based on the existing error annotation, we generated three quality scores. Accuracy and fluency score were calculated from content and language error frequencies, respectively. Overall translation score (tq) was obtained using raw overall error statistics. All scores were normalised to the ST word count.

In an attempt to generate more theoretically justified scores, we also produced a weighted version of these scores using an elaborate weighting scheme. *Weighted* accuracy and fluency scores in Table 6.7 take into account severities of errors (e.g. a critical error weights more than a minor error). Additionally, these scores are weighted by individual categories of content and language errors as recommended by *TAUS*. For example, an error in collocation has a greater weight than a spelling error. The overall quality score uses the severities and annotated positive assessment (kudos) in its weighted variant. This explains why tq score can go over the maximum bound of the absolute scale (100).

Scaled tq score is tq score re-calculated for a 0-100 scale, using Formula 6.1:

$$scaledscore = \frac{100 \times (score - min)}{max - min} \tag{6.1}$$

where *score* is normalised error frequency, *min* and *max* are respective values of the unscaled score range.

The scaling was done to put the scores from error annotation on the same measuring scale as DA scores.

Detailed account of converting error statistics to scores is given in Section 4.4.2 (see page 129).

| | Min | Max | Mean (+/-STD) |
|---|---|---|---|
| Unweighted scores | | | |
| accuracy | 93.55 | 100 | 98.12 +/-1.04 |
| fluency | 90.88 | 99.8 | 97.36 +/-1.32 |
| tq | 87.62 | 99.56 | 95.49 +/-1.92 |
| scaled_tq | 0 | 100 | 65.86 +/-16.04 |
| Weighted scores | | | |
| accuracy | 61.28 | 98.9 | 86.03 +/-6.69 |
| fluency | 59.21 | 98.46 | 86.52 +/-6.11 |
| tq | 72.67 | 100.36 | 92.69 +/-4.35 |
| scaled_tq | 0 | 100 | 72.3 +/-15.72 |

Table 6.7: Quantitative parameters of error-based scores across 553 documents

Table 6.7 has the parameters of each quality score distribution (across 553 translations). Note that weighted scores are better distributed across the 0-100 scale, while unweighted scores are hurdled around the maximum value.

Table 6.8 has the results of the document-level regression tasks for the error-based scores across seven representation.

| | accuracy | | fluency | | **tq** | | scaled_tq | |
|---|---|---|---|---|---|---|---|---|
| | Unweighted scores | | | | | | | |
| | $r$ | RMSE | $r$ | RMSE | $r$ | RMSE | $r$ | RMSE |
| **UD** | **0.43** | 0.95 | **0.43** | 1.18 | **0.45** | 1.72 | 0.32 | 15.5 |
| all | 0.41 | 0.96 | 0.43 | 1.19 | 0.43 | 1.74 | 0.32 | 15.56 |
| quest61 | 0.37 | 0.98 | 0.42 | 1.16 | 0.36 | 1.73 | 0.31 | 15.22 |
| tfidf | 0.48 | 0.92 | 0.49 | 1.14 | 0.47 | 1.69 | 0.39 | 15.59 |
| TQmono-m | 0.46 | 0.94 | 0.47 | 1.15 | 0.45 | 1.68 | 0.35 | 15.3 |
| ruRoberta | 0.51 | 0.91 | 0.53 | 1.08 | 0.54 | 1.57 | 0.44 | 14.84 |
| mdeberta3 | 0.58 | 0.87 | 0.58 | 1.05 | 0.62 | 1.5 | 0.56 | 14.62 |
| | Weighted scores | | | | | | | |
| UD | 0.38 | 6.31 | 0.37 | 5.72 | 0.44 | 4.02 | **0.38** | 15.13 |
| all | 0.38 | 6.34 | 0.35 | 5.77 | 0.43 | 4.06 | 0.36 | 15.18 |
| quest61 | 0.32 | 6.34 | 0.33 | 5.68 | 0.37 | 4.1 | 0.33 | 15.13 |
| tfidf | 0.4 | 6.36 | 0.39 | 5.78 | 0.48 | 4.04 | 0.45 | 15.23 |
| TQmono-m | 0.4 | 6.13 | 0.37 | 5.64 | 0.47 | 3.92 | 0.38 | 14.85 |
| ruRoberta | 0.48 | 5.98 | 0.45 | 5.47 | 0.57 | 3.74 | 0.5 | 14.53 |
| **mdeberta3** | **0.62** | 5.76 | **0.59** | 5.21 | **0.67** | 3.54 | **0.63** | 14.19 |

Table 6.8: Regression results for error-based quality scores (553 documents)

**Performance of UD against other representation** Similar to classification experiments on quality labels, morphosyntactic features proposed in this thesis were significantly outperformed only by *mdeberta3* on unweighted accuracy and tq scores, and by *mdeberta3* and *ruRoberta* on fluency.

The best result of $r = 0.67$, not significantly outperformed in any other setting, was seen for *mdeberta3* for weighted overall quality (cf. $r = 0.45$ for UD features).

As expected, *collgram* and *ngram* features implicitly present in *all* were not helpful and introduced noise, showing slightly lower performance than UD feature set across all scores.

Interestingly, a custom *tf-idf* representation[1] consistently performed better than both morphosyntactic features and TQ-estimation-dedicated cross-lingual sentence embedding model from *TransQuest* in absolute terms, but we do not have enough data to see statistically significant results on these comparisons.

[1]produced on lemmatised corpus with proper names and numbers substituted with respective placeholders, see page 142

**Comparison of regressors across types of scores** From Table 6.8, Spearman's $r$ was higher for overall quality score (tq) than for accuracy, fluency or scaled_tq for both weighted and unweighted versions. However, most differences between the results on accuracy, fluency or tq were not statistically significant. Whenever there was a significant difference (mostly between weighted fluency and tq), the results were always better for tq.

Significant differences between weighted and unweighted versions were seen only for fluency on *ruRoberta* and for scaled_tq on *mdeberta3*. However, it is impossible to say which version of scores was more learnable because the unweighted score returned higher results in the former setting and the weighted score was superior in the latter. It is noteworthy that weighted scores returned much larger error between the gold scores and predictions as measured by RMSE. Based on these observations, unweighted scores are preferable to weighted scores.

**Scaled version of the overall quality scores was inferior to unscaled version** on almost all representations. Scaling the quality score tripled the average prediction error, too.

In the subsequent experiments reported in this chapter, weighted scores and scaled version of overall quality (scaled_tq) are excluded from the analysis.

These results draw attention to a curious finding. Contrary to the expectations, we did not see statistically significant differences between accuracy and fluency in terms of performance on any representations, including translationese features. If these scores captured independent of each other aspects of quality, we would expect higher results for fluency on our TL-centred representations (even if not on translationese-aware UD features). We can cautiously suggest that **our data did not contain translations that were very inaccurate but fluent, and vice versa**. The two aspects of quality seem to be correlated. Indeed, on weighted scores Spearman's correlation coefficient between them was 0.897[2]. This might mean that **an HTQE approach to aimed at fluency can be a useful proxy for overall quality in human translation quality estimation based on document-level quality scores from error annotation**. Another explanation is lack of training data. Our experiments were run on 553 annotated translations and

[2]We argue that weighted scores, which take into account the type of errors and their severities reflect the actual opinion of assessors' on quality

did not reveal the differences between the two aspects of quality as captures by error annotation aggregated at document level. If more data is needed to see the differences between the two, then the difference must be really small. A more plausible explanation is that **scores from error annotation do not reflect translationese properties of translations, at least when they are generated for entire documents**.

**Feature analysis** Feature selection for UD set did not yield any sizeable improvements in performance on either weighted or unweighted scores. On weighted scores, the maximum gain was for accuracy, where the value for correlation coefficient went up from 0.38 to 0.42 on a selection of 29 features. On unweighted scores, a selection of 44 UD features increased the regression result on accuracy only by 0.01 (from 0.43 to 0.44).

Limiting QuEst++ feature set to the best-performing 21 features resulted in a very modest gain of 0.03 on accuracy. No selection of features was able to improve the performance on fluency and tq scores for either UD or QuEst++.

To reveal the features that were better correlated with various quality scores, we used each of them as a single input to a *linear regression* algorithm as implemented in *scikit-learn*, with the library default parameters.

Univariate correlation analysis did not reveal any features from UD, collgram, ngram or QuEst++ feature sets that were associated with any quality score for more than Spearman's $r = 0.274$. Despite the correlation between the proposed quality indicators and various quality scores was small, we noticed that some important quality predictors from UD features (from the top 10 better-correlated features) contributed to all scores while others remain specific to accuracy or fluency. The intersection of top 10 features by correlation with unweighted accuracy, fluency and tq included *nsubj, acl:relcl, obj*. These relatively strong predictors shared by all error-annotated scores capture syntactic complexity of sentence in translations: number of explicit syntactic subject and object dependencies across all clauses (nsubj, obj and number of relative clauses (acl:relcl). All of them were weakly but positively correlated with the scores. This brings us to a counter-intuitive conclusion that the more well-developed finite clauses there were in the sentence, the higher was the quality score. This also contradicts the results on binary labels (see page 6.1.2).

One way to interpret this contradiction is to suggest that holistic document-level

perception of quality incorporated translationese-related properties of translations, while error-based quality scores, even from fluency errors, did not reflect translation solutions that can be explained by translationese. The fact that the more prominent quality predictors were equally associated with accuracy, fluency and tq, together with the observation that this association was weak ($r < 0.243$), confirms that even if accuracy and fluency reflected different aspects of translation quality, these aspects were not related to translationese.

The look at other translationese features that were found among the top 10 by correlation with either accuracy or fluency suggests an interpretation coherent with the translationese hypothesis of quality for some of them. For example, accuracy scores had returned $r = 0.226$ for Spearman's correlation with the frequency of negative particles. This aligns well with the translationese explanation: the more negative particles, the higher the translation quality. Less fluent translations tend to underuse negative constructions and follow the English affirmative way of expression. Accuracy scores also had $r = -0.1$ correlation with epistemic markers. The overuse of Russian literal translations for the phrases like *certainly, as far as we know, of course, no doubt, it appears that*, which are more common in English, seemed to predict lower-ranking translations. The most prominent negative indicator of fluency was the frequency of modal predicates ($r = -0.106$). From translationese theory perspective, in the English-to-Russian translation, it makes sense that the overuse of modal predicates imposed by SL should signal lower quality. If these observations are acceptable, given their low potency, they do not explain either error-based scores or the difference between them convincingly.

The comparable **performance of QuEst++ for all error-based scores was based on ST complexity features**. For the QuEst++ feature set the intersection between best-performing features on accuracy and fluency included six out of 10 top features. Five of them were complexity (ST-based) features that *measured average number of translations per source word* counted with various Giza probability thresholds (including #1016, #1018, #1020, #1022, #1024, e.g. document-level QuEst++ feature #1024: "average number of translations per source word in the sentence (threshold in giza1: prob $> 0.05$)"). Apart from that, number of sentences in the target document (#9801) was relatively well correlated with all scores. It indicated that the more sentences there were in a translation, the higher was the quality score.

The error-annotated dataset included multiple translations for 47 English sources.

Apparently, the differences between them were most contributive to the regressor performance. We looked through the extended lists of top 15 most-associated features for accuracy and fluency, and found that they represented the properties of ST (with the exception of two adequacy features). Over-reliance of this classifier on the ST side fits well with no distinction between accuracy and fluency. Indeed, a more demanding ST can trigger both content and language errors, as we demonstrated in (Kunilovskaya et al., 2022). However, it also shows that this feature set does not pick up any relevant distinctions between translations per se.

Analysis of intersections in *tf-idf* top features did not reveal any patterns and was not enlightening.

### 6.2.3 Document Level: Direct Assessment

While our annotation setup for perceived quality was sentence-level, document-level DA scores were obtained by averaging sentence scores. Recall that the annotation setup did not present sentences in isolation. Instead, the raters could see the entire parallel document. They were asked to read the ST before annotation.

In the description below, *da_mean* stands for the mean raw score between the raters. Following the practice adopted in WMT quality estimation shared tasks, raw scores were standardised by rater using z-transformation and then averaged to obtain *da_zmean* scores.

Also, as explained in Section 4.4.3, we were able to use only two ratings from six sentence-wise cross-annotations of 150 documents due to strict thresholds for internal rater reliability. Ten documents were discarded because their annotations were incomplete or there were other technical problems preventing the recovery of trustworthy results.

The level of agreement between the two selected ratings in terms of Spearman's rank correlation co-efficient for document-level scores was 0.503 ($p < 0.05$).

Table 6.9 has the Spearman's correlation of averaged DA score (da_mean) and the three major scores from error annotation. All results are significant at the confidence level of 5%.

| | unweighted | | | weighted | | |
|---|---|---|---|---|---|---|
| | accuracy | fluency | tq | accuracy | fluency | tq |
| da_mean | 0.509 | 0.169 | 0.391 | 0.514 | 0.408 | 0.559 |

Table 6.9: Correlation of averaged DA scores with quality aspects scores from error annotation (Spearman's r)

It can be seen from Table 6.9 that weighted error-based scores were better aligned with perceived quality. To an extent, it justifies the error weighting scheme based on error categories and severities. Accuracy scores and overall scores are more correlated with perceived quality than fluency. It means that the **raters penalised content errors more than language errors**.

Performance of UD against other representation

| | Min | Max | Mean +/-STD |
|---|---|---|---|
| da_mean | 56.36 | 95.57 | 83.65 +/-7.99 |
| da_zmean | -3.10 | 1.34 | 0.00 +/-0.88 |

Table 6.10: Document-level: Parameters of DA scores distribution

The resulting distribution of scores from DA annotation for 140 documents (averaged from sentence scores) has the parameters reflected in Table 6.10. They are also visualised in Figure 4.10 on page 4.4.3. The distribution of DA scores is skewed to the right.

According to the results, presented in Table 6.11, there was no consistency in performance of the regressors across the representations on either raw mean DA scores or on mean z-transformed scores. In fact, none of the differences between the two $r$ columns in Table 6.11 were statistically significant.

Further on, there were no statistically significant differences between the representations shown in rows, either. Despite we can observe that the value for Spearman's $r$, averaged across 10 folds of the experiment, was higher for *mdeberta3*, the performance of the regressor was very volatile from fold to fold, ranging from $r = 0.068$ to $r = 0.758$.

| | da_mean | | da_zmean | |
|---|---|---|---|---|
| rep | $r$ | RMSE | $r$ | RMSE |
| UD | **0.23** | 7.27 | 0.16 | 0.81 |
| all | 0.21 | 7.29 | **0.23** | 0.79 |
| quest61 | 0.18 | 7.44 | 0.11 | 0.88 |
| tfidf | 0.19 | 7.56 | 0.2 | 0.79 |
| TQmono-m | 0.23 | 7.31 | 0.22 | 0.78 |
| ruRoberta | 0.22 | 7.35 | 0.23 | 0.78 |
| **mdeberta3** | **0.37** | 7.22 | **0.32** | 0.75 |

Table 6.11: Document-level: Regression results for DA scores

We have to conclude that none of the representations was more successful than the other in learning DA scores. Clearly, this is the effect of small training data (140 documents) in conjunction with no strong signal coming from any attempted representations.

The results for document-level experiments on DA scores were twice lower than on error-based scores reported in Section 6.2.2 above. To rule out the impact of data size (553 vs 140 documents for error-based and DA scores, respectively), we obtained the error-based results on 140 documents cross-annotated for errors and DA (see Table 6.12).

| | accuracy | | **fluency** | | tq | | scaled_tq | |
|---|---|---|---|---|---|---|---|---|
| | Unweighted scores | | | | | | | |
| | $r$ | RMSE | $r$ | RMSE | $r$ | RMSE | $r$ | RMSE |
| **UD** | **0.13** | 1.32 | **0.4** | 1.48 | **0.24** | 2.48 | **0.21** | 20.62 |
| all | 0.02 | 1.34 | 0.33 | 1.51 | 0.14 | 2.5 | 0.13 | 20.66 |
| quest61 | -0.07 | 1.46 | 0.39 | 1.47 | 0.15 | 2.47 | 0.1 | 20.59 |
| tfidf | 0.07 | 1.3 | 0.45 | 1.48 | 0.22 | 2.42 | 0.21 | 20.59 |
| TQmono-m | 0.01 | 1.38 | 0.42 | 1.46 | 0.24 | 2.42 | 0.2 | 20.54 |
| ruRoberta | 0.04 | 1.38 | 0.45 | 1.41 | 0.25 | 2.37 | 0.29 | 20.31 |
| mdeberta3 | 0.52 | 1.12 | **0.54** | 1.35 | 0.52 | 2.1 | 0.46 | 19.93 |

Table 6.12: Regression results for 140 cross-annotated documents

*Unweighted fluency* was the only error-based score that remained robust on a four times smaller dataset (553 documents vs. 140 documents cross-annotated for errors and DA). The regression results for unweighted fluency were almost twice higher than for DA scores on the same dataset, and the differences were statistically significant, unlike most other differences between Tables 6.11 and 6.12.

It seems that this setup returned the expected higher learnability of fluency, as opposed to accuracy, on the representations limited to the TL-side of the parallel corpora. Besides, it showed that error-based scores seem to be more consistent and reliable than DA scores.

However, the distributions of accuracy scores in the error-annotated and cross-annotated datasets (see Figure 6.4 were quite different. In the full dataset it is close to normal, while in the cross-annotated it is almost binomial. The fluency and tq scores were not as affected by the selection of 140 documents. Recall from Section 4.4.3 (page 4.4.3) that the documents for cross annotation were specifically selected to be in the intersection of error-annotated data and binary quality data. We also aimed to include approximately the same number of *good* and *bad* translations. The observed change in distribution of accuracy scores provides a further confirmation that human judges paid more attention to the accuracy of meaning transfer than to disfluencies in translations.

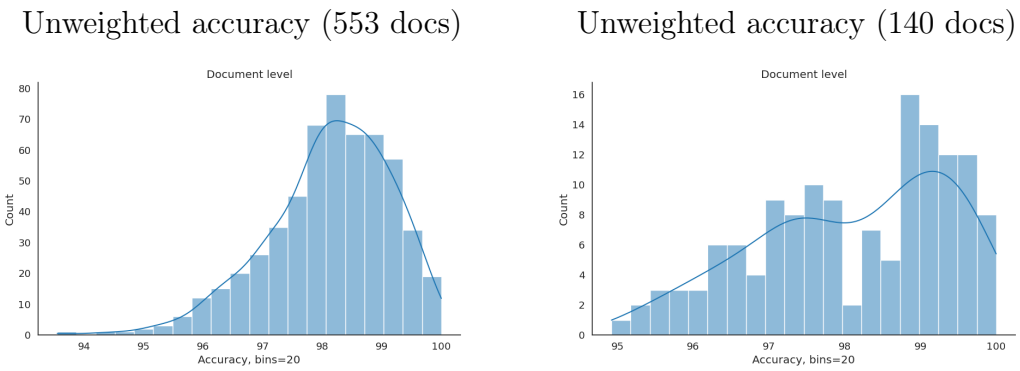



Figure 6.4: Distributions of accuracy scores for error-annotated and cross-annotated datasets

These details on data selection invalidate the differences in performance between accuracy and fluency on the cross-annotated dataset, while the observation about lack of difference between the two aspects from the point of view of all attempted representations stands. This conclusion was confirmed by **lack of statistical difference between accuracy and fluency** for all attempted representations on a randomly selected 140 error-annotated documents. The results of this experiment are aggregated in Table 6.13.

| | accuracy | | fluency | | tq | | scaled_tq | |
|---|---|---|---|---|---|---|---|---|
| | Unweighted scores | | | | | | | |
| | $r$ | RMSE | $r$ | RMSE | $r$ | RMSE | $r$ | RMSE |
| **UD** | 0.2 | 1.11 | **0.36** | 1.02 | **0.36** | 1.72 | **0.34** | 14.8 |
| all | 0.28 | 1.08 | 0.25 | 1.04 | 0.3 | 1.74 | 0.27 | 14.9 |
| quest64 | **0.29** | 1.09 | 0.29 | 1.03 | 0.28 | 1.77 | 0.26 | 14.94 |
| tfidf | 0.38 | 1.06 | 0.33 | 1.03 | 0.33 | 1.74 | 0.3 | 15.04 |
| TQmono-m | 0.32 | 1.05 | 0.31 | 1.04 | 0.34 | 1.7 | 0.23 | 14.8 |
| ruRoberta-large | 0.38 | 1.06 | 0.34 | 1.01 | 0.38 | 1.69 | 0.39 | 14.74 |
| **mdeberta3-base** | **0.53** | 0.99 | **0.45** | 0.95 | **0.58** | 1.56 | 0.54 | 14.6 |

Table 6.13: Regression results for random 140 documents with error-based scores

With regard to the **comparison between error-based and DA scores**, we have seen statistically significant results only for *mdeberta3*, which returned higher results for

accuracy, tq and scaled_tq scores than for da_zmean scores. All other comparisons were inconclusive.

Interestingly, *mdeberta3* was exceptional among all representations, including other embedding models. It returned stable and relatively high results regardless the type of error-based score, size of the dataset and distribution of scores.

In sum, the experiments on an random selection of error-annotated documents confirmed **no difference between the accuracy and fluency on any representations** reported in Section 6.2.2 (see Table 6.8). We also **failed to obtain convincing proof that either error-based or DA scores were better aligned with the properties of translation** as represented in this work, except for *mdeberta3* vectors that performed better on accuracy, tq and scaled_tq scores than on da_zmean scores.

**Feature analysis** We continue to trace the features that can be more predictive of the scores than other features from UD and QuEst++ using feature selection and univariate tests, and limiting the analysis to the top ten most successful quality predictors by both or either of these methods.

In this paragraph we compared the set of features most associated with error-based scores to the set of features most correlated with DA scores.

Feature selection improved the results for *da_mean* on UD from 0.23 to 0.27 by limiting the feature set to just *deverbals, nmod, nsubj*. The results on QuEst++ went from 0.18 to 0.24 after limiting the feature set to 24 features. However, we forgo their analysis because the differences were not statistically significant. The analysis of individual correlation of feature values and scores revealed **a noticeably different set of features correlated with DA scores than with error-based scores for UD and QuEst++**. In fact, there was only one UD feature that was in top ten by correlation with DA scores and with three error-based scores (*nsubj*), while *mark, neg, advcl* and *sentlength* were shared with only accuracy or fluency, respectively. None of these quality indicators support the translationese approach to quality, except, maybe, number of negations (*neg*). If anything, these results mean that the longer and more complex the sentence, the higher the quality. In all likelihood, in **DA assessment setup translationese properties of texts were ignored**.

UD features that were correlated with DA scores (but not with error-based scores)

included *mdd, aux, tempseq, aux:pass, mark, copula, advers*. They were positively correlated with DA scores. Higher MDD means higher comprehension difficulty, with sentences requiring additional effort for transforming longer linear distances into syntactic trees at understanding time (Jing and Liu, 2015, p.162). The value in the DA dataset was greater than in non-translated Russian. Increased frequencies of copula and auxiliary verbs, analytical passives are signs of more noticeable translationese. The frequency of adversative discourse markers was negatively correlated with DA scores ($r = -0.224$), i.e. the more explicit markers of contrast were used in a translation, the lower was the perceived quality (e.g. *однако, несмотря на, тем не менее* [however, despite, nonetheless]). These observations contradict our assumption that higher-quality translations should show less signs of translationese.

One common quality indicator among 61 QuEst++ features that cut across all types of quality scores (both DA and error-based) was *the ratio of verbs in the source and target (#1093)*. It was negatively correlated with the quality, i.e. the fewer verbs in the target, the higher the ratio, and the lower the quality. If anything, this observation contradicts the expectations in the context of English-Russian language pair and translationese tendencies as was the case with some other results for translationese features for error-based (see page 6.2.2) and DA scores on UD features.

Other DA predictors also useful for accuracy and tq (#1052, #1085), or fluency (#1004) were:

**1052** average bigram frequency in quartile 3 of frequency (lower frequency words) in the source corpus,

**1085** ratio of percentage of content words in the source and target,

**1004** number of tokens in the target over number of tokens in the source.

They were positively correlated with the respective quality scores.

To give examples of features that were specific for predicting DA scores, we list the most correlative variables below (in the decreasing order based on individual correlation with *da_mean* score):

**9992** ratio of word repetitions in target and source documents,

**9993** ratio of lemma repetition between target and source documents,

**1003** ratio of number of tokens in source and target,

**1004** number of tokens in the target over number of tokens in the source,

**1052** average bigram frequency in quartile 3 of frequency (lower frequency words) in the source corpus).

Interestingly, this list primarily includes adequacy features. Recall that error-based scores had ST complexity parameters as major quality predictors.

Lack of intersection between error-based quality predictors and DA quality predictors from two interpretable feature sets make us conclude that these **types of assessment reflect dissimilar aspects of translation quality**.

Admittedly, this feature analysis has very limited reliability, given weak association of any individual features with the DA scores (maximum $r = 0.274$ for negative particles from UD feature set) and the general poor performance of any feature set on DA scores. However, the interpretation of most correlative features indicates that **DA method does not reflect translationese properties of document**.

### 6.2.4 Sentence Level: Error-based and DA Scores

Given that error-based and DA scores originated at sentence level, we extended these experiments of the sentence-level version of the dataset. It included 3,224 sentences from 140 documents with error-based and DA scores. At sentence level the distribution of scores was not as affected by the selection of a polarised smaller sample and remained heavily right-skewed for all scores as shown in Figure 4.8 (page 133). Nonetheless, we report results on randomly selected 3,224 sentences. The stack of experiments was run only for unweighted scores. For consistency we present the results on the full dataset spanning 12,369 sentences, too.

| | accuracy | | fluency | | tq | | scaled_tq | |
|---|---|---|---|---|---|---|---|---|
| Error-based scores (12,369 sentences) | | | | | | | | |
| | r | RMSE | r | RMSE | r | RMSE | r | RMSE |
| UD | 0.21 | 3.91 | 0.24 | 4.7 | 0.21 | 3.91 | 0.2 | 11.97 |
| all | 0.21 | 3.91 | 0.26 | 4.67 | 0.21 | 3.91 | 0.21 | 11.97 |
| **quest70** | **0.24** | 3.93 | **0.26** | 4.67 | **0.24** | 3.93 | **0.23** | 11.96 |
| tfidf | 0.31 | 3.8 | 0.28 | 4.62 | 0.31 | 3.8 | 0.31 | 11.84 |
| TQmono-m | 0.27 | 3.88 | 0.27 | 4.66 | 0.27 | 3.88 | 0.26 | 11.94 |
| **ruRoberta** | **0.35** | 3.78 | 0.35 | 4.52 | **0.35** | 3.78 | **0.33** | 11.74 |
| mdeberta3 | 0.32 | 3.82 | **0.36** | 4.49 | 0.32 | 3.82 | 0.31 | 11.8 |
| Error-based scores (random 3,224 sentences) | | | | | | | | |
| UD | 0.17 | 3.84 | 0.23 | 4.74 | 0.17 | 3.84 | 0.17 | 11.77 |
| all | 0.17 | 3.84 | 0.24 | 4.71 | 0.17 | 3.84 | 0.17 | 11.77 |
| quest70 | 0.14 | 3.87 | 0.25 | 4.7 | 0.14 | 3.87 | 0.13 | 11.79 |
| tfidf | 0.21 | 3.77 | 0.22 | 4.73 | 0.21 | 3.77 | 0.21 | 11.69 |
| TQmono-m | 0.21 | 3.83 | 0.24 | 4.72 | 0.21 | 3.83 | 0.21 | 11.75 |
| **ruRoberta** | **0.29** | 3.75 | 0.31 | 4.6 | **0.29** | 3.75 | **0.29** | 11.63 |
| mdeberta3 | 0.27 | 3.77 | **0.33** | 4.56 | 0.27 | 3.77 | 0.27 | 11.66 |

Table 6.14: Sentence-level results for unweighted error-based scores

At sentence level there was enough data (on 12,369 sentences) to establish that **translationese-aware UD features were relatively competitive only for fluency scores**. On the full dataset, they were outperformed for this type of quality only by *ruRoberta* and *mdeberta3*. For accuracy and tq, translationese-aware feature set was inferior to all other representations, including *quest70*, *tf-idf* and *TQmono-m* in addition to *ruRoberta* and *mdeberta3*. This inferiority was lost on 3,224 sentences, except for *ruRoberta* and *mdeberta3*. It means that the differences in performance between UD, on the one hand, and *quest70*, *tf-idf* and *TQmono-m*, on the other hand, were small.

The results for (unweighted) error-based scores at sentence level were generally almost twice lower than at document level for the full dataset of 553 documents/12,369 sentences (see Unweighted scores in Table 6.8). For example, for fluency UD features achieved $r = 0.43$ at document level and $r = 0.24$ at sentence level, while *mdeberta3*'s results were $r = 0.58$ and $r = 0.36$, respectively. Manual frequency features were expected to be affected by sparsity. However, it is not clear why a dedicated sentence-level QuEst++ feature set and vectorised representations did not benefit from the shift from document to sentence level.

We noticed that the drop in performance caused by limiting the data to a (random) forth of the original size was bigger for documents than for sentences. For example, the ratio of $r$ obtained from full/small document-level experiments for unweighted fluency on UD features was 1.19, while for sentence level it was 1.04. For quest70 these ratios were 1.44 vs 1.04, for *ruRoberta* 1.56 vs 1.12 and for *mdeberta3* 1.2 vs 1.09. This means that there was less volatility between the folds, and the learning process was more stable at sentence level, even if not as successful as at document level, when enough data is available. These observations across many representations make us hypothesise that **error-based quality scores reflected the properties of texts better than they fit the properties of sentences**.

Sentence-level experiments, contrary to document-level, *yielded evidence that fluency scores were more learnable than accuracy scores*, at least, on UD and *mdeberta3* for both full and reduced datasets. This is an expected and comforting outcome which confirms the sanity of the error annotation and of the distinctions between the two aspects of quality in principle when enough data is available. At the same time, for the practical purposes of human translation quality estimation, these distinctions are negligibly small.

To cross-examine our conclusion from document level that neither error-based nor DA scores were better aligned with the properties of translation than the other, we obtained the results for DA at sentence level.

Spearman's $r$ between rater1 and rater2 on sentence-level DA scores was 0.408 ($p < 0.05$). This was a bit lower than for document-level perspective ($r = 0.503$).

To give more details on the parameters of the cross-annotated sentence-level variant of the dataset, we report correlation between error-based scores and DA in Table 6.15, measured with Spearman's $r$.

| | unweighted | | | weighted | | |
|---|---|---|---|---|---|---|
| | accuracy | fluency | tq | accuracy | fluency | tq |
| da_mean | 0.291 | 0.149 | 0.291 | 0.311 | 0.165 | 0.33 |

Table 6.15: Sentence-level: Correlation of averaged DA scores with error-based scores

Correlation between two assessment methods at sentence-level was much lower than at document level (cf. Table 6.9, page 199).

Table 6.16 presents the results for DA scores on sentence-level version of the dataset.

| | da_mean | | da_zmean | |
|---|---|---|---|---|
| rep | r | RMSE | r | RMSE |
| 2raters | | | | |
| UD | 0.29 | 12.89 | 0.29 | 0.84 |
| all | 0.3 | 12.86 | 0.3 | 0.83 |
| **quest70** | **0.33** | 12.74 | **0.33** | 0.82 |
| tfidf | 0.27 | 12.94 | 0.31 | 0.82 |
| TQmono-m | 0.35 | 12.56 | 0.37 | 0.81 |
| **ruRoberta** | **0.39** | 12.41 | **0.42** | 0.78 |
| mdeberta3 | 0.39 | 12.45 | 0.38 | 0.80 |

Table 6.16: Sentence-level: Regression results for DA scores

With regard to the main question of this thesis – whether translationese-aware UD features were competitive against other representations in predicting quality – the results for DA scores demonstrated that they were inferior to *ruRoberta* and *mdeberta3* as before, but also to *TQmono-m*. Note that *TQmono-m* was specifically trained for predicting this type of quality – DA scores – on MT data.

Sentence-level results for DA scores were better than at document-level on all representations, but especially for QuEst++ (cf. da_mean: $r = 0.33$ vs $r = 0.18$) and *ruRoberta* ($r = 0.39$

vs $r = 0.22$). Curiously, this was not the case with error-based scores where we discussed a unanimous downward slide on all representations (see page 206). This emphasises the differences between the two manual assessment methods. **Error annotation (as implemented in the RusLTC, see Section 4.4.2, page 125) is more fit to produce document-level scores than DA setup**.

At document level, there was no evidence that either error-based scores or DA scores returned higher results on any representations, except marginal superiority for accuracy and tq over DA on *mdeberta3* (see page 202).

Sentence level experiments provide ample evidence that **DA scores were more learnable than any error-based scores, given the same sample size**. In fact, the experiments on sentences from the same 140 cross-annotated documents and on a random sample of 3224 sentences returned results that were higher for DA scores than for error-based scores on an overwhelming majority of representations. For example, UD results was $r = 0.29$ for da_mean score and $r = 0.23$ for fluency on random 3224 sentences; *mdeberta3* returned $r = 0.39$ for da_mean score and $r = 0.33$ for fluency.

We leave out feature analysis at sentence level for two reasons. First, we believe that translationese is primarily a document-level property. Second, UD features are heavily affected by sparsity in this setting.

## 6.3 Discussion

This chapter evaluates the ability of linguistically-motivated translationese indicators to predict professional varieties and human translation quality against alternative document representations and using various annotation methods to generate quality judgments.

The performance of ML algorithms on hand-engineered translationese features was contrasted with the results on document-level features from QuEst++, a MT quality estimation framework, on *tf-idf* representation, and on vectors from two types of embeddings: cross-lingual sentence embeddings and word embedding (large dedicate TL model and the most recent multilingual SOTA in representation learning).

We evaluated the ability of our representations to distinguish between student and professional translations viewed in this work as a possible proxy for translationese-related quality.

In quality estimation tasks, we explored the learnability of three types of quality judgments, including (i) binary quality labels from holistic document-level assessment of translations, (ii) continuous quality scores generated from error annotation and (iii) continuous quality scores from sentence-level DA annotation.

**Translationese indicators for translation varieties and 'good/bad' translations**
The classification experiments to distinguish professional and student translations demonstrated that the differences between them were only marginally related to translationese. The week signal picked by translationese indicators was overshadowed by other (semantic and functional) dissimilarities between the two parallel corpora. This limitation of our setup was anticipated, and measures were taken to obtain most comparable subsets from a much larger corpus resource (see Section 3.2.2, page 77).

We found that in this experiment QuEst++ significantly outperformed all other hand-engineered feature sets, reaching $F1 = 86.17$ after feature selection. The success of the feature set was explained by the performance of complexity features that captured the properties of the STs. Our representations did not target accuracy specifically. The only feature set that implicitly compared ST and TT to access relations between them was QuEst++. It did not perform well on any task targeting quality labels or scores. Therefore, we omitted feature analysis for those tasks. On the varieties dataset, QuEst++ features performed well but it relied on ST-related features that in our setting cannot be reliably interpreted as quality indicators. The three best-performing QuEst++ adequacy features demonstrated that student translations were less wordy but more repetitive than professional translations (see page 181). QuEst++ fluency features confirmed that student translations were more perplexity to a TL model. They probably used a greater ratio of n-grams unseen in non-translations. n student and professional translations without using ST was limited to several fluency features and three adequacy features. The findings on these features confirmed that student translations were more repetitive and more perplexing to a trigram LM, i.e. more distinct from non-translations than professional translations.

The best-performing UD subset of 15 features achieved $F1 = 79.39$ on this task. This subset included only six features that were also strong translationese predictors, mostly reflecting the syntactic complexity of target sentences (number of clauses, clausal

complements and subordinating conjunctions). On these features student translations were closer to the reference corpus although the effect size of the differences was small. On other translationese features, contributing to the varieties distinction, students seem to demonstrate a bit stronger translationese tendencies due to underuse of negation, overuse of copula verbs and relative clauses but we did not have enough data to obtain statistical evidence for this observation.

The most prominent quality indicators from UD features did not have strong association with the labels, maximising at the level of 0.34 for *parataxis*. Comparative analysis for this feature demonstrated that students preferred to use explicit markers of cohesion, which led to a lack of asyndetically introduced elements. This makes student translations most distinct from professional translations, and especially from non-translations. **Translation quality classification** proved to be an even more challenging task for all representations. First, there was little content variation between the categories because good and bad translations were produced from the same 57 English STs. Second, the quality labelled dataset was almost twice smaller than professionalism dataset (708 vs 360 instances).

Despite these limitations, UD features were much more successful in this task, demonstrating that the **amount of translationese is well-correlated with the document-level quality assessment**.

Linguistically-motivated translationese-aware features were inferior only to *ruRoberta* and *mdeberta3*, large pre-trained word embedding models. Unlike in the other experiments, RFECV feature selection offered a significant gain in performance, bringing our classification result on UD features to $F1 = 68.9$ (vs $F1 = 78.14$ on *mdeberta3*).

We argue that the reasonably good performance of out-of-the-box contextualised embeddings on this dataset, although not enlightening from the point of view of translation studies, indicates that human assessment of quality at document-level is actually related to the objective properties of the texts, and might not be as subjective as it is common to believe, at least when it is trusted to translation professionals.

The results of feature analysis suggest that quality distinctions were much more related to translationese than differences between professional varieties of translation. We demonstrated that there were only three (out of 26) quality predictors that were not among translationese indicators. Professionalism turned out to be a more vague concept than quality, at least from the point of view of translationese indicators.

An explanatory analysis of translational trends in *bad* translations demonstrated that **lower quality is almost entirely associated with the lack of ability to recognise and counteract the ST influence**. For the features where there was enough evidence to establish differences between *good* and *bad* translations, we observed translationese trends which implied passively following ST patterns, with little or no effort to adapt them to the TL norm. These included

Generally, lower-ranking translations were found further away from non-translations in Russian than higher-ranking translations for the same STs across most analysed features.

**Translationese indicators and continuous quality scores** In Sections 6.2.2, 6.2.3 and 6.2.4, we attempted to predict quality scores generated from error annotation and in DA setup. Our focus remains on evaluating the performance of translationese indicators against other representations, and on explaining the findings using interpretable features.

Importantly, scores from error-annotation allow to disintegrate translation quality into accuracy and fluency, and test whether fluency returns higher results on the representations that have access to the TL-side of the corpus only. Besides, scores offer an opportunity to extend document-level experiments to sentence level. On the one hand, this allows to alleviate the problem of small training data at document-level. On the other hand, it helps uncover important properties of scores generated by each assessment method.

Section 6.2 brought together several aspects of this multi-faceted comparison, namely, experimental results are compared across:

- seven representations, focused on comparative performance of UD features,
- various scores and their versions from two manual assessment methods, including on a cross-annotated subset,
- most successful features from hand-engineered sets.

We have seen that the outcomes of these comparisons were largely dependent on the level of analysis: document or sentence.

**Translationese-aware features vs. contextualised embeddings** Our major hypothesis was that translation quality (calculated from error statistics or perceived in DA setting) can be captured by looking at the amount of translationese in a translation.

Cross-lingual, dedicated Russian and multi-lingual embedding models (*TQmono-m, ruRoberta, mdeberta3*, respectively) were comparable to the proposed hand-engineered translationese-aware feature sets in that they did not recur to the source text and encoded the properties of the target side of the dataset only.

Table 6.17 summarises the outcomes of all experiments showing the models that were statistically superior to UD. It uses unweighted versions of error-based scores and results on random 140/3224 error-annotated instances.

| | Document level | | Sentence level | |
|---|---|---|---|---|
| | 553 docs | 140 docs | 12,369 sents | 3,224 sents |
| accuracy | mdeberta3 | mdeberta3 | all reps | ruRoberta<br>mdeberta3 |
| fluency | ruRoberta<br>mdeberta3 | – | ruRoberta<br>mdeberta3 | ruRoberta<br>mdeberta3 |
| tq | mdeberta3 | – | all reps | ruRoberta<br>mdeberta3 |
| da_mean | NA | – | NA | TQmono-m<br>ruRoberta<br>mdeberta3 |
| da_zmean | NA | – | NA | TQmono-m<br>ruRoberta<br>mdeberta3 |

Table 6.17: Representation superior to UD on quality scores

**For document-level scores, linguistically-motivated translationese features returned compatible performance against other representations**. As can be seen from Table 6.17, they were statistically surpassed only by *mdeberta3*, the most recent embedding approach and current SOTA in NLP representation learning, in all experiments on both error-based and DA scores. Admittedly, the performance on document-level DA scores and on random 140 error-annotated documents was very poor due to the size

of the dataset, and no statistically significant differences were seen between any two representation. Lack of a bigger annotated data prevents us from attempting any fine-tuning methods, especially for DA-annotated dataset.

**At sentence level** on all available error-annotated instances (12,369), **UD features performed well on fluency scores only**, being secondary to the two word embedding models. For accuracy and overall translation quality, UD features were not competitive. This outcome make sense: if accuracy reflected semantic similarity between ST and TT, translationese indicators should not be able to predict this type of quality. However, these differences in performance are very small and can be picked on a large annotated dataset. They go away if not enough data is available (see last column in Table 6.17).

For sentence-level DA scores translationese-aware features were inferior to *ruRoberta* and *mdeberta3* as before, but also to *TQmono-m*, a model optimised to predict DA scores on MT data. This observation might mean that **DA scores have specific properties shared between MT dataset and HT dataset** built for the purposes of this research.

Although direct comparisons with any previous work on document-level automatic human translation quality estimation is problematic, we can compare our results to those reported by (Yuan, 2018). They reported Pearson's $r = 0.72$ against *QuEst++* baseline result of $r = 0.67$ for the best-performing subset of their manual features on document-level English-to-Chinese dataset based on overall quality (for more details see our description of their work on page 102). We obtained the highest Spearman's $r = 0.45$ on UD features for overall quality, with QuEst++ features returning $r = 0.36$ (the difference is not statistically significant). The absolute best result on our dataset is $r = 0.67$ (*mdeberta3*). This comparison demonstrates that in relative terms our features offered a comparable increase on the QuEst++ baseline for our dataset.

**Error-based scores vs. DA** Our experiments put under scrutiny 10 continuous quality scores and their versions from two annotation methods. In general, it seems that no modification of raw scores was able to yield better results across all representations than their raw counterparts. If anything they performed worse. This applies to weighted error-based scores, scaled overall quality from error annotation (scaled_tq) and to z-transformed DA scores (da_mean).

Another important comparative conclusion about various scores is that **error-based**

**scores seem to capture document properties of documents better than DA scores, but DA scores are superior for sentence-level experiments**. Although in both annotation scenarios annotators were exposed to the entire documents, while providing judgments on the properties of individual sentences, error-based scores were more learnable than DA scores at document level than at sentence level (including on the dataset of the same size). We have shown that unlike DA scores, error-based scores demonstrated a considerable decrease in performance when moving from document level to sentence level on the same dataset (see pages 206 and 6.2.4). Document-level nature of error-based scores is further supported by a bigger drop in their performance for documents than for sentences when moving from full dataset (553 documents/12,369 sentences) to random forth of the data (140 documents/3224 sentences). The sentence-level performance of error-based scores was poor at the full dataset, so it did not suffer as great a tumble on limited data. Finally, we made an interesting observation about the relations between accuracy and fluency, the two major theoretical aspects of quality, which stir a lot of controversy in the literature on benchmarking translation quality by humans. According to our results, the difference between accuracy and fluency was visible only at sentence level and only for some representations. In particular, only UD and *mdeberta3* returned expected statistically higher results for fluency than for accuracy on both full and reduced datasets. At document level, there was no evidence that either accuracy or fluency was more aligned with the properties of documents picked by any attempted representations. This lack of evidence at document level can be explained by the lack of training data. However, our experiments were run on 553 annotated translations. If more data is needed to see the differences between the two, then the difference must be really small. Another explanation is that error-based judgments capture local issues in translations and should not be used to generate overall quality scores, especially across all sentences in a document. Finally, we can hypothesise that, in line with the assumption made in this thesis, accuracy errors indeed introduce disfluencies in human translation, and therefore, scores based on content errors can be learnt using representations that do not have access to ST (such as UD which achieved $r = 0.427$ and *ruRoberta* with $r = 0.534$ at document-level). With regard to UD representation, especially given its relatively low result, another plausible explanation is that **scores from error annotation do not reflect translationese properties of translations,**

**at least when they are generated for entire documents**. It is quite possible that errors were focused on other, more localised, imperfections of student translations, instead of translationese. It stands to reason, because translationese is in essence a 'distributed' statistical property of translation which is difficult to put a finger on in individual sentence in error annotation. A more general interpretation of this outcome, which also incorporated *mdeberta3* performance, can be that the **distinction between accuracy and fluency is negligibly small for all practical purposes of quality estimation at document level**.

The relation between scores from the two assessment methods provide an important insight on the nature of human perception of translation quality in DA setting. We showed that (i) accuracy was almost four times more correlated with DA scores than fluency; (ii) weights increased this correlation and (iii) weighted overall quality returned the highest correlation with DA scores, which reached $r = 0.559$ (see Table 6.9). Professional human raters in both annotation schemes laid greater emphasis on accuracy. At the same time, most attempted representations, including those that were not focused on translationese-related fluency and could pick up the signal coming from the ST directly (QuEst++) or indirectly (*TQmono-m*, *mdeberta3*), had a tendency to return higher results for fluency or tq, especially on smaller datasets. It means that machines struggle with picking those aspects of quality that are deemed more important by the humans. Arguably, fluency is the least 'non-binary' aspect of translation quality. Unlike accuracy, it does not rely too much on human understanding and interpretation of language, context or cross-linguistic correspondences.

**Feature analysis results by score type** In the regression experiments, the feature selection process and univariate feature analysis proved less enlightening than in classification tasks. It was limited to the document-level scores only.

We have seen but very modest improvements in regressor performance, if any, on feature selections returned by RFECV (setup with a linear kernel, $C = 1.0$ and Spearman's r as a scoring function).

Similar to (Yuan, 2018), we found that the strongest quality indicators were well-associated with all error-based quality scores (three most correlated UD features and top six QuEst++ features out of 10). This supports the opinion that accuracy and fluency

are but artificial constructs which are difficult to tease apart objectively.

We also observed that the features most correlated error-based and DA scores were quite different. There was one 'universal' quality indicator in the selections of 10 most effective UD and QuEst++ features that was also well-associated with the other type of quality scores. They were number of nominal subjects and *ratio of verbs in the source and target* (1093), negatively-correlated with all quality scores.

We can hypothesise that error-based and DA scores captured dissimilar parameters of translation quality. A bit larger intersection between good DA and good accuracy indicators confirms that raters who were tasked with perceived quality annotation gave more weight to distortions in semantic relations between ST and TT than to imperfections in linguistic expression.

With regard to the relation between the amount of translationese and quality scores, we have to conclude that **error annotation and, especially DA, largely ignore translationese properties of translations**. At document level, where the relation between translationese and quality was expected, scores from these assessment methods returned correlations with translationese indicators that partially contradicted the expectations based on translationese perspective of quality. Contrary to what we observed on binary quality labels from holistic assessment of translations (see discussion on page 6.3), feature analysis in experiments with quality scores did not reveal a consistent link between translationese properties of documents and their quality. If anything, the interpretation of many linguistic features from UD and QuEst++ feature sets demonstrated that more translationese was correlated with higher quality, especially for DA scores (see analysis on pages 6.2.2 and 6.2.3).

# Chapter 7 Conclusion

This thesis applies translationese theory to the study of human translation quality. Using machine learning methods, it explores the extent to which translationese properties of text can explain the quality judgments obtained by holistic assessment of translated documents, error annotation and direct assessment of sentences in context. Additionally, we use professional and student translations as ontological varieties that can be related to quality.

The investigation has an explanatory focus. It employs linguistically-motivated translationese-aware features to discern patterns in translational behaviour and individual types of translation solutions that can be associated with lack of professionalism or lower translation quality. Our findings contribute to translationese studies and are applicable in translation education. Besides, the thesis offers a comprehensive comparative study of translation quality assessment methods, including those used to benchmark translation quality in machine translation. It sheds light on the important properties of quality labels and scores originating from various approaches to manual quality assessment.

To prove the hypothesis that translationese properties are related to translation quality, we need to show that the features useful in predicting translationese can also be used to predict translation quality and/or that translationese-related phenomena are more seen in lower quality translations. For example, Hu and Kübler (2021, p.357) conclude their multilingual translationese study by stating: "translations tend to have more relative clauses. This seems to be truly universal despite obvious differences in the source languages". If they are right, normalised frequency of relative clauses in a translated document should be among strong translation predictors for English-Russian language pair. Further on, if our hypothesis stands, relative clauses should be on the list of strong quality predictors and/or by relatively well correlated with the quality label or score.

The relation between translationese and quality has been suggested in the literature many times (see Section 2.2). However, we are not aware of research that would put this proposition to an empirical test by machine learning methods, at least, for human translations. This thesis frames the problem as a human translation quality estimation

task supported by statistical testing and feature analysis. Our major focus is on document-level quality, which is known to be under-researched in MTQE, including due to lack of annotated data and limitations of representation learning. Importantly, out hypothesis makes us rely on predicting quality using the target language side of the parallel corpora, while ignoring the relation between ST and TT. We contend that in human translations produced in good faith, deviations from semantic similarity between ST and TT would usually distort the natural flow of text and create weird nonsensical renditions suggestive of accuracy errors.

The aim and the methodology of this research place it at the interface of a number of disciplines. It builds on the knowledge and findings in

- translation studies, especially in the areas of translationese studies, TQA, professional competence research and pedagogy;
- corpus linguistics, especially building learner and parallel corpora and corpus analysis methods;
- MTQE, and applications to HTQE;
- ML and NLP for computational methodology.

**Contributions** Our contributions can be summarised as follows:

1. *We designed and rigorously tested three sets of linguistically-motivated translationese indicators for English-Russian translations of mass-media texts.*

   They include 60 morphosyntactic features extracted from Universal Dependencies annotation and two smaller sets of abstract lexical features, including 24 collocational features and 11 features from n-gram language models. The selection process was guided by translationese theory as well as by findings from previous empirical studies. In particular, these features include most translationese indicators from a comprehensive set developed by Volansky, Ordan and Wintner (2015), which is used as a baseline in many translationese studies. With regard to the UD features, care was taken to include only cross-linguistically comparable isomorphic features informed by the contrastive analysis of English and Russian that could hypothetically pick specificity of language choices in the situation of translation (Section 2.3).

Our experimental results on two independent collections of translations proved that structural morphosyntactic features were by far more effective in distinguishing translations and non-translations. They were also very competitive against vector representations from pre-trained contextualised word embedding models, including a dedicated Russian language model and a multilingual *mdeberta3*, the current SOTA in generic representation learning (Section 5.2).

2. *We identify most prominent translationese indicators for English-Russian mass-media translation, associated with 'shining-through' trend in translational behaviour, using bottom-up approach.*

   Multivariate analysis through feature selection revealed a large intersection between sets of translationese features that were most effective in predicting student or professional translations against non-translations. In univariate analysis, these translation varieties were shown to exhibit similar trends in translational behaviour along the same dimensions of analysis, most notably, 'shining-through'.

   The most prominent translationese indicators were among syntactic, rather than morphological, features, and signalled a tendency to produce longer and more complex sentences in translation. We confirmed some of the translational distortions anticipated by translation theory and practical textbooks on English-to-Russian translation. Namely, frequencies of the following items tend to be higher in translations than in non-translations: additive discourse markers, analytical passives, copula verbs, modal predicates, personal pronouns, finite verbs and determiners. The explorations at the lexical level indicated that translations were likely to have more of unusual strings, unseen in a large TL model. They relied on a smaller variety of collocations, and the collocation in translations were less familiar to the language model (Section 5.3).

3. *We obtained evidence that professionalism-related differences between translations were much less related to translationese than distinctions between good and bad translations.*

   Translationese-aware UD feature set was outperformed on this task by QuEst++ features. Despite QuEst++ success can be largely explained by differences in the STs, findings on prominent fluency and adequacy features demonstrated that student translations were more perplexing to a LM trained on non-translations, most likely

due to 'strange strings' and awkward word choice. Student translations were also found more repetitive. Analysis of a few strong translationese indicators, selected as predictive of professionalism, returned contradictory results. Some expected translational behaviours were actually more noticeable in professional translations, while other – in student translations. For example, professional translations were found to contain longer sentences, more epistemic markers and modal predicates than student translations. However, on other translationese features, professionals demonstrated more ability to recognise and counteract translationese (more indirect objects and negative particles, fewer analytical passives). Besides, the association between translationese features and labels was very low, with many strong translationese predictors being irrelevant for professionalism opposition (Section 6.1).

4. *We provided evidence that prominent translationese indicators were indeed predictive of document-level quality labels from holistic assessment confirming that the amount of translationese was well-associated with human perception of translation quality.*

   Linguistically-motivated morphosyntactic translationese features were competitive against other representations on document-level quality labels and scores. They outperformed QuEst++ and were second only large pre-trained word embedding models. Feature analysis provided further overwhelming support for our hypothesis that translationese indicators can be useful in predicting quality. Note that this finding is limited to holistic assessment of document-level quality. We demonstrated that, in the experiment on binary quality labels, lower-ranking translations were always more distant from the expected target language norm and were well-associated with stronger translationese trends (Section 6.1.2)

5. *We demonstrated that quality judgments from three assessment methods revealed dissimilar properties: they accounted for translationese to various extent and performed better at document or sentence level.*

   Quality scores derived from error annotation and sentence-level direct assessment scores were shown to be less translationese-aware than document-level holistic perception. The study of association between translationese indicators and error-based scores returned ambiguous results. Some results were coherent with translationese approach to quality (the higher the score, the less translationese), while for some features

higher-scoring translations manifested more translationese. Comparative performance of various representation demonstrated that error-based scores returned higher results at document level, while scores from direct assessment performed better at sentence level. Additionally, in our setting there was no evidence that any attempted document-level representations returned better results for accuracy or fluency. Given that some of our representations did not have direct or indirect access to the SL, this outcome raises doubts about the practical reliability of this distinction at least at the document level (Sections 6.1.2 and 6.2).

6. *We release relatively large document- and sentence-level aligned datasets based on student translations with holistic document-level labels, error-based and DA scores that can be used for HTQE.* The datasets contain English-to-Russian translations of mass-media texts from the *RusLTC* and were produced by senior students majoring in translation at BA or MA (specialist) levels. Documents with binary labels (360 documents) include several top and several bottom translations to 57 English texts that were offered to students as test assignments or translation competition tasks. Submitted translations were sorted by the final assessment result (agreed between jurors/examiners) that was returned to the student or announced at a competition. Error-based scores are frequencies of translation errors normalised to the document word count. The annotations were produced as part of translation education process across a number of years for multiple student translations to 46 English sources. This dataset counts 553 documents/12,369 sentences. Direct assessment experiment involved final-year BA students on a translation or linguistics degree programme. The dataset includes two most internally consistent ratings for 140 documents (3,224 sentences). This subset is in intersection with binary labels dataset and with error-annotated dataset. Detailed description of the datasets are presented in Section 4.4.

**Summary of the chapters** This paragraph offers a brief overview of the essential component of this thesis presented in chapters. Each of Chapters 2- 4 is a unity of theoretical and practical part, which describes a separate stage in the project. These chapters open with conceptual foundations and empirical evidence from previous work to explain practical implementations described in the last section of each chapter.

**Chapter 2** introduces the theoretical underpinning of this research rooted in translationese

studies. It explains the approach pursued in this work in the context of previous and recent developments in translationese theory. Section 2.2 explains the hypothesised link between translationese and translation quality. The final section of Chapter 2 offers a detailed description of the proposed feature sets, including their selection and extraction, motivated by the research goal, our stance on translationese and adopted methodology.

We defined translationese as any statistical deviations of translations from comparable non-translations, regardless of their origin and nature. Following the overwhelming evidence from previous research, presented in Section 2.1.1, our approach assumes that shining-through and normalisation are major translationese trends.

The overview of most prominent methodologies in translationese studies demonstrates how theoretical foundations and research goals are linked to feature engineering. We characterise corpus-based and computational methodologies that can rely on surface, linguistic or linguistically-motivated features, depending on the amount of external resources and knowledge is necessary to extract the features. The choice of theoretical approach and major research designs in translationese studies are presented as following top-down or bottom-up approaches. Researchers can pursue linguistic-theoretical and explanatory goals or be after most effective engineering solutions for a related NLP task. Along these dimensions, this research can be described as an explanatory bottom-up approach based on computational methods and linguistically-motivated features.

Section 2.2 brings together theoretical and empirical evidence that translationese can be predictive of translation quality and levels of professional expertise. Most studies that put this hypothesis to a test were corpus-based and traced a few individual hand-picked features (like sentence length or adverbial quantifiers). There were a few computational studies that explored this relation in the context of MT. In this section, we discussed the feasibility of a quality estimation approach to human translation that relies only on the properties of translations and effectively ignores accuracy as an aspect of quality. This approach is assumed in a translationese-based HTQE. Briefly the major pro-arguments include reported evidence for the lack of distinction between accuracy and fluency in annotation outcomes, greater subjective impact of disfluencies in translation perception, including leading to comprehension difficulties, findings from comparative analysis of professional and student translations and error analysis.

Finally, Section 2.3 describes three feature sets proposed in this thesis. We hypothesised

that they should be effective in predicting translations and translation quality. We proposed a set of 60 structural morphosyntactic features that included normalised frequencies of default syntactic relations, morphological forms and categories as annotated in UD framework as well as features that relied on more sophisticated extraction algorithms such as mean hierarchical distance or cumulative frequencies of listed causative discourse markers, for example. The key principles that guided the selection of structural features included the following considerations. They were contend independent, cross-linguistically comparable (hence, language pair specific), motivated by observations from translation practice textbooks, corpus-based error analysis, previous research in translationese studies. We also heeded interpretability and reliability of extraction. In an attempt to pick translationese phenomena that can be manifested at lexical level we proposed two sets of abstract lexical features that were not based on surface strings directly. They reflected collocational and distributional properties of translations.

Chapter 3 discusses previous work related to learner translator corpora, the study of professionalism in translation and building register-comparable corpora for translationese studies. Its major aim is to present the textual data used in this research and to provide motivations for corpus building solutions.

Our discussion is focused on issues of using student data annotated in real-life contexts as research data and on findings about differences between professional and student translations. Student and professional translations reveal a lot of similarities between them, and researcher usually find less differences than expected. There were attempts to interpret differences between students and professional from the point of view of translationese theory, including relating these distinctions to quality.

Importantly, Section 3.1.3 discuses the issue of datasets and resources comparability in translationese studies. We explore the concept of comparability in NLP and in corpus-based research. The section shows that the more important aspect of comparability in translationese studies in related to register to use a term which reflects text-internal approach to functional text categorisation. Functional text comparability of parallel corpora and reference corpus is key in translationese studies, because registers are known to trigger different types of translationese and professional translational norms are register-specific. Building register-comparable corpora is treated as an important preliminary step in this research.

Section 3.2.1 has the description of data collection. As student translations are our major type of text and they come from our own corpus project, we shared our experience with an emphasis on technological issues in learner translator corpus-building. In our experience they are associated with handling multi-parallel data and adding error annotation to a parallel corpus. The section presents the rationale behind selecting various subsets of student translations from the RusLTC aimed at obtaining homogeneous collections for each experiment. In this section we refer to previous research on *Functional Text Dimensions* framework that was used increase comparability of subcorpora in this research. These efforts included processing a English sources from professional and student parallel corpora in a monolingual setup: we calculated the most comparable subsets of professional and student parallel corpora, and a cross-lingual task aimed at finding a subset of Russian non-translations most similar to the joint collection of English sources. Finally, this thesis relies on large language collections of mass-media texts in Englshand Russian, used for training LMs. These resources are also introduced in Section 3.2.1.

Chapter 4 introduces key theoretical concepts related to translation quality, and discusses existing practical approaches to manual assessment and automatic quality estimation. Importantly, it contains the description of annotation experiments and reliability studies associated with the quality judgments used as learning targets in this research.

Throughout the chapter we maintained a comparative focus bringing together approaches in TS and in MT with regard to quality. We demonstrated that TS typically describes quality in terms of three aspects, with adequacy having priority over accuracy and fluency. MT researches express doubt that these distinctions are practical from the point of view of reliable annotation, and if quality is disaggregated into aspects, accuracy and fluency are distinguished.

Another point of divergence between HT and MT is the level of quality judgments. In TS word-level of sentence-level quality are difficult notions, while most research in MT quality estimation/evaluation is done on these levels, the typical use cases of MT. Research at document-level is also held back by lack of annotated resources and limitations of existing approaches in representation learning.

Section 4.2 describes translation quality estimation as a computational linguistics and NLP task in MT, including best-performing approaches in MTQE and a few existing cross-overs to HTQE. Understandably, the potential of feature-learning approaches and neural

networks is being actively explored in the field, but explainability remains a bottleneck.

Section 4.3 has a detailed description of best practices in quality assessment: focused on grading in HT and on producing quality benchmarks for ML experiments in MT. It covers such competing methods as rubrics and error annotation (in HT), and error annotation, direct assessment, post-editing (in MT). We notice that there is more pessimism in MT with regard to reliability of assessment, especially at document level, which is not upheld in our research. Also, MT has more awareness that various assessment methods capture dissimilar properties of translations, an property of quality assessment that was clearly revealed in this thesis.

Finally, Section 4.4 explained how the quality labels and scores used in this thesis were obtained and demonstrated their reliability. Our labels were based on the status of translators with regard to professionalism (students vs professionals) and on the outcomes of document-level holistic assessment, which was limited to contrasting categories of 'good' and 'bad' translations of the same ST. Quality scores were obtained from error-annotation and sentence-level direct assessment. The agreement between university translation teachers and final-year under-graduate students (DA experiment) ranged between 0.467 and 0.734 in various settings, with the lowest result seen for sentence-level DA by students.

Chapter 5 presents the results of translation detection tasks based on professional and student translations. We refer to these experiments as translationese classifications, in which we test the ability of the proposed feature sets to distinguish translations and non-translations (i.e. register-comparable documents originally authored in the TL; also known to as reference corpus in translationese studies). A considerable part of this chapter is given to feature analysis and description of translational properties of classification categories, facilitated by the use of interpretable explicit features.

This section provides details of experimental setup used in four classification tasks reported in this thesis. Our major learning algorithm was SVM with linear kernel. For sanity check we also implemented a neural classifier which returned very similar results. All classification tasks use a similar set of representations alternative to the proposed feature sets. They put the proposed features in the perspective of the achievable in each task. Besides, their comparative analysis of their performance facilitated the interpretations of various properties of documents in each category. Alternative representations were selected to offer a spectrum of representation approaches ranging from *tf-idf* to

multilingual *mdeberts3*, the current SOTA in word-level representation learning. We also included cross-lingual sentence embedding models trained on STS task and on DA scores in a MTQE task. While all documents in our classifications are in Russian, we added a large dedicated Russian word-embedding model.

Translationese classifications on both student and professional translations return high results in all settings. It is not surprising as most representations are content-dependent and capture topical differences between translations and non-translations. It is fairly easy to predict translation using strings of proper names only. Delexicalised theoretically-motivated morphosyntactic features proved to be very competitive against those very strong alternatives, while abstract lexical features, capturing collocational and distributional properties of documents did not live up to our expectations.

The comparison of outcomes for student and professional translations allowed to reveal strong translationese indicators and establish specificity of each translation variety. Professional translations were characterised as being more homogeneous and predictable, demonstrating fairly stable language use than student translations. Despite our efforts to obtain functionally comparable collections, we established that English texts translated by professionals might have been structurally more complex, which was a confounding factor in our experiments.

Feature analysis was based on RFECV, a feature selection algorithm that determined a best-performing subset of most contributing features. It helped to identify strong translationese features, which were additionally verified by statistical analysis. We have seen that features included in multivariate patterns by feature selection would often have reasonably high correlation with the labels in a single-feature classifier setting.

Establishing good translationese predictors does not allow to describe properties of translations by itself. To understand actual trends in translational behaviour and to prove or disprove the theoretical expectations that guided feature engineering, we performed three-way comparative analysis, looking at frequencies of features in sources, targets and translations. Depending on the outcomes of these comparisons we were able to characterise features as contributing to *anglicisation, shining-through, overuse of ST items, underuse of TL items, normalisation, adaptations*. They cover a continuum of translational strategies defined by the amount of active attempts aimed at reconciling the differences between SL and TL.

Our research confirmed that carrying over SL frequency patterns into translations, whenever possible, was the major factor explaining translational deviations from the expected TL norm. Recall that we tried to keep our features cross-linguistically comparable and more or less open too alternative solutions, and not rigidly imposed by the language system. There were more translationese indicators associated with this translational behaviour than with the solutions which involved selecting restructuring original sentences and using non-isomorphic Russian items. SL/TL-independent translationese was more often seen in student translations. For example, the used more additive discourse markers than were seen in their STs or than was expected in the TL. One overwhelmingly strong trend seen on several syntactical features that cut across both translation collection was sentence lengthening and increased syntactic complexity.

Although abstract lexical features were not as successful as UD features in the translation detection task, their analysis revealed some additional properties of translations. For example, student translations were found more perplexing to an n-gram language model, which might suggest reduced fluency.

Chapter 6 is the main experimental chapter, which brings together classification experiments on professionalism and quality labels, and regression experiments on document- and sentence-level quality scores.

In this chapter the performance of translationese indicators is contrasted with the performance of QuEst++ features, a well-known set for MTQE.

Although professional and student translations were well-distinguishable on translationese-aware features, the major differences between them were not associated with translationese. Feature analysis demonstrated that professional translators might be even more prone to translationese than students along some dimensions. Higher predictability of professional translations can be partly explained by their higher homogeneity arising from more stable behaviour than that in student translations. However, our results are less reliable than expected, because professional and student parallel corpora demonstrated some differences in the STs that were interfering with the interpretation.

The experiments with binary quality labels, which reflected document-level assessment based on holistic perception of student translations, yielded support for our hypothesis that lower-ranking translations should exhibit more translationese than higher-ranking translations. UD features achieved comparatively good results in this task, returning F1

of 0.68 on the best-performing feature subset. This result was secondary only to *mdeberta3* vectors with $F1 = 0.78$. Feature analysis demonstrated that strong translationese indicators were indeed among the strongest quality predictors, and that lower-ranking translations were more distant from the expected TL norm than higher-ranking translations.

The experiments on quality scores were less convincing in terms of performance, but they were enlightening about the properties of error-annotation and sentence-level Direct Assessment as methods of generating quality scores.

Comparative analysis of the regression results on a number of representations and feature analysis indicated that error annotation reflected translationese properties of documents to a greater extent than DA scores. But this is probably because error-based scores are essentially document-level metrics. Recall that translationese is also by nature a statistical property, which is more visible at document level. The performance of error-based scores at sentence level was twice lower than at document level, which was not the case for DA scores. Feature analysis results for error-based scores were at best ambiguous: for some features they were in line with translationese hypothesis of quality, but for others they contradicted the main assumption.

Feature analysis for DA assessment scores had a tendency to contradict the expectation that more translationese means lower quality. Many well-expressed translationese features positively correlated with DA scores. It means that this type of assessment was ignorant of translationese phenomena. Probably, it is not surprising, as DA scores returned better results than error-based score at sentence-level and were shown to be much better correlated with accuracy score than with fluency scores.

In our experiments, there was **no evidence that either accuracy or fluency scores were better aligned with the properties of texts** as represented by seven various methods, including those that were not designed to pick semantic inaccuracies due to no access to the SL side of the corpus. However, at sentence level the difference in performance on accuracy and fluency was seen for UD, QuEst++ and *mdeberta3* even for a small dataset of 3,224 random sentences. This invites a conclusion that error-based judgments capture local issues in translations and should not be used to generate overall quality scores, especially across all sentences in a document. At the same time, in our setup we cannot rule out the possibility that accuracy errors do introduce disfluencies in human translation, and therefore, scores based on content errors can be learnt using

representations that do not have access to ST (such as *ruRoberta*). Essentially, this line of interpretation negates the difference between aspects of quality for any experiments that are not aimed at word-level quality.

With regard to the **differences between results for error-based and DA scores**, they were lower for any error-based score than for DA (da_mean) on all representations, given a fair random subset of 3,224 sentences (cf. lower part of Table 6.14 and Table 6.16). We interpreted this difference as reflecting the particularities of the annotation setup. Unlike error annotation, raters involved in DA were asked to pass judgment on every sentence. In error-annotated corpus 58% of sentences had at least one span annotated. This also explains lower correlation between error-based and DA scores at sentence level than at document level. For example, for document-level unweighted accuracy and mean DA score: $r = 0.509$ and for fluency: $r = 0.169$, while at sentence level those correlations were 0.291 and 0.149, respectively. The correlation between quality scores from two manual assessment methods indicates that in direct assessment setup humans tend to give preference to accuracy over fluency. Moreover, a suggested weighting scheme based on TAUS recommendations, improved the correlation between error-annotated and perceived quality. It justifies the perceived difference between various error categories and between severity levels.

Concerning the performance of translationese indicators on various types of quality judgments, we can conclude that each of them captured a specific aspect of quality, which may or may not be related to translationese. Holistic document-level perception manifested in binary quality labels was shown to be the most translationese-aware quality judgment. Lower-ranking (*bad*) translations clearly demonstrated expected translationese tendencies (see page 190). Findings from feature analysis in experiments with error-based document-scores indicated that this type of judgment might be focused on other aspects of translation quality, which did not reflect translationese properties. Note that these conclusions were only possible as a result of feature analysis on hand-crafted feature sets. The overall classification results were not enlightening as to which properties of translations were picked by the classifiers and regressors.

**Limitations and Future work** This research has a number of limitations. They can be arranged around the major steps of this research. Some limitations are suggestive of

future work.

The most obvious and painful limitations are have to do with feature design and extraction. We limited ourselves only to those theoretically known translationese phenomena that we were able to extract with relative reliability. However, the proposed features can be ambiguous in terms of interpretation. For example, the use of any types of pronouns, especially possessive and indefinite/total is expected to be different under the pressure of the ST. Although our analysis failed to find empiric proof for that, we will be weary of dismissing pronouns as a translationese give-away. It is likely that this feature requires taking into account some context which was ignored in our approach. A more effective feature set might take into account specific structures which are expected to be translationese-prone.

The proposed approach to feature extraction does not link ST and translation items. Although we rely on document- and sentence-level aligned corpora, there is no reliable way to say that specific ST items triggered the overuse of isomorphic structures in translations. It can be argued that this connection is useful only for detailed descriptions of translators' solutions, and not for overall characteristics of translated language. The same challenging English items can push up frequencies of various Russian items depending on which coping strategy is preferred by a translator. For example, English infinitives functioning as adverbial modifiers of purpose (structures like *to do X, Y did Z*) do not have a parallel structure in Russian. Typical solutions involve (i) a infinitive clauses necessarily introduced by a marker of inferential relations (e.g. *чтобы противостоять этой реакции, Кремль мобилизовал как вещательные, так и онлайн-СМИ*), (ii) a prepositional phrase with a deverbal noun (e.g. *для формирования общественного мнения в интернете Кремль создал мощную цифровую инфраструктуру из ботов и платных авторов*) or (iii) a predicative complement construction with a verb of purpose (e.g. *Кремль пытается противостоять этой реакции путем мобилизации троллей*). These solutions have various degrees of similarity to the original structure (with the first one being most literal) and drive up counts for various translationese features.

Nonetheless, developing more context-aware bilingual features might be more beneficial for interpretation. An attempt to contextualise translationese features was made in (Eetemadi and Toutanova, 2015) and developed by Sominsky and Wintner (2019). They proposed leveraging word-level alignments to produce a vocabulary of lexically anchored minimal

translation units inspired by phrase-based machine translation techniques. A simpler approach based on human translations is described in Hu, Li and Kübler (2018), who used constituent parse trees and dependency triples as translationese features.

It is well-known that one of the major syntactic problem of human English-to-Russian translations is word order. However, our features do not focus specific translationese-prone structures, such as time and place adverbials and prepositional phrases in the absolute sentence end. Frequencies of bigrams, even PoS bigrams are not specific enough to capture these phenomena.

Our feature analysis was by design limited to capture translationese associated with SL/TL contrast. However, deviant frequencies of items in translations can be generated by other factors as well. For example, it is unlikely that increase sentence length and sentence complexity can be entirely explained by the languages involved. We agree that this trend might be related to coping with cognitive load and lack of paraphrasing skills and flexibility, i.e. by deficiency in productive aspects of TL competence.

With regard to the data collection, we have shown that the outcomes of translationese classifiers on student and professional translations, and experiments on professionalism were affected by the differences in the underlying source text collections. This limitation was to an extent remedied by feature analysis, but calls for a special treatment of translations based on diverging ST collections. For example, Popovic (2020) proposed to minimise the effects of diverging ST collections by using *relative difference* between ST and TT as values for TT features.

In terms of experimental setup, we envision several possible extensions to further explore the relation between HTQE and translationese properties of translations. They include ideas that had to be left outside the scopes of this research due to considerations of space.

A straightforward approach to test the relevance of translationese properties to quality prediction is proposed by Aharoni (2015). They used translationese classifier probabilities for translated class as a single input to a quality predictor on the assumption that the higher the translationese classification accuracy, the lower the expected translation quality. In future we want to try merging UD features and embeddings to try and improve the results on quality labels. Current results make us hypothesise that manual features and embeddings capture different quality-related aspects of translated documents.

An important insight from our analysis is that human perception, at least at sentence level, emphasise accuracy. It is obvious that a well-rounded effective approach to automatic quality estimation should te able to measure similarity between ST and TT. We hope to improve our previous attempts to complement translationese approach with ST-TT semantic similarity module, and develop a effective NLP solution for document-level human translation quality estimation.

Finally, our error-annotated data is big enough – it is the size of the WMT datasets, at least – to test neural solutions for word-level quality estimation task existing in MT and to adapt them to predict error spans in human translations. This direction of future work might be forestalled by the challenges of converting *brat* error annotations into a reliable ML dataset.

In sum, holistic human assessment was reasonably well-aligned with the amount of translationese, especially in its shining-through form, while error-based and DA-annotated quality captured other properties of translations that were unrelated to manifestations of translationese. The distinctions between professional varieties were not rooted in translationese, either.

We noticed that error-based scores for accuracy and fluency did not reveal the expected difference in performance on representations that, by design, were able to capture the TL properties only. This confirmed our intuition that in human translation, **inaccurate translations are usually also disfluent**. Hence, a reliable approach to capture fluency can be a useful proxy for overall quality in human translation quality estimation.

# References


Aharoni, R. (2015), Automatic detection of machine translated text and translation quality estimation, Master's thesis, Bar-Ilan University. Available at: www.roeeaharoni.com/thesis.pdf

Aharoni, R., Koppel, M. and Goldberg, Y. (2014), 'Automatic detection of machine translated text and translation quality estimation', *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, Baltimore, 23–25 June, pp. 289–295, doi: 10.3115/v1/P14-2048.

Ahrenberg, L. (2017), 'Comparing machine translation and human translation: A case study', *Proceedings of the First Workshop on Human-Informed Translation and Interpreting Technology (HiT-IT)*, Varna, 7 September, pp. 21–28, doi: 10.26615/978-954-452-042-7_003.

Akhbardeh, F., Arkhangorodsky, A., Biesialska, M., Bojar, O., Chatterjee, R., Chaudhary, V., Costa-jussa, M. R., Espa{\˜n}a-Bonet, C., Fan, A. and Federmann, C. (2021), 'Findings of the 2021 Conference on Machine Translation (WMT21)', *6th Conference on Machine Translation, WMT 2021 - Proceedings* pp. 1–88, doi: 10.18653/v1/w19-5301.

Alfuraih, R. F. (2019), 'The undergraduate learner translator corpus: a new resource for translation studies and computational linguistics', *Language Resources and Evaluation* 54(3), pp. 1–30, doi: 10.1007/s10579-019-09472-6.

Aranberri, N. (2020), 'Can translationese features help users select an MT system for post-editing?', *Procesamiento de Lenguaje Natural* 64(1), pp. 93–100, doi: 10.26342/2020-64-11.

Artstein, R. and Poesio, M. (2008), 'Inter-Coder Agreement for Computational Linguistics', *Computational Linguistics* 34(4), pp. 555–596, doi: 10.1162/coli.07-034-R2.

Baker, M. (1993), 'Corpus linguistics and translation studies: Implications and applications', in Baker, M., Francis, G. and Tognini-Bonelli, E. (eds.) *Text and Technology: In honour of John Sinclair*, Amsterdam: John Benjamins Publishing Company, pp. 232–250, doi: 10.1075/z.64.15bak.

Baker, M. (1996), 'Corpus-based translation studies: the challenges that lie ahead', *Terminology, LSP and translation: Studies in language engineering, in honour of Juan C. Sager*. pp. 175–186, doi: 10.1075/btl.18.17bak.

Baker, M. (2011), *In other words: a coursebook on translation*, 2 ed., London and New York: Routledge.

Baroni, M. and Bernardini, S. (2006), 'A new approach to the study of translationese: Machine-learning the difference between original and translated text', *Literary and Linguistic Computing* 21(3), pp. 259–274, doi: 10.1093/llc/fqi039.

Becher, V. (2011), Explicitation and implicitation in translation. A corpus-based study of English-German and German-English translations of business texts, PhD thesis, University of Hamburg. Available at: ediss.sub.uni-hamburg.de/handle/ediss/4186

Beltagy, I., Peters, M. E. and Cohan, A. (2020), 'Longformer: The long-document transformer', *arXiv preprint arXiv:2004.05150* .

Belyaev, M., Popova, E. and Serova, I. (2010), *Grammar and Syntax in Newspaper Translation*, Saint Petersburg: Saint Petersburg State University of Economics. Available at: http://gen.lib.rus.ec/book/index.php?md5=22975A055EBD9BE97B1130776F4EE8A7

Berman, A. (1985), 'Translation and the trials of the foreign (translated by Lawrence Venuti)', in Venuti, L. and Baker, M. (eds.) *The Translations Studies Reader (2000)*, London/New York: Routledge, pp. 284–297.

Bernardini, S. and Baroni, M. (2005), 'Spotting translationese: A corpus-driven approach using support vector machines', *Proceedings of Corpus Linguistics 2005*, Vol. 1, University of Birmingham, pp. 1–12. Available at: tinyurl.com/yk5z3sck. (Accessed: 17 June 2022).

Bestgen, Y. (2017), 'Beyond single-word measures: L2 writing assessment, lexical richness and formulaic competence', *System* pp. 65–78, doi: 10.1016/j.system.2017.08.004.

Bestgen, Y. and Granger, S. (2014), 'Quantifying the development of phraseological competence in L2 English writing: An automated approach', *Journal of Second Language Writing* 26, pp. 28–41, doi: 10.1016/j.jslw.2014.09.004.

Biber, D. (1988), *Variation across speech and writing*, Cambridge: Cambridge University Press, doi: 10.1017/CBO9780511621024.

Biber, D. (1995), *Dimensions of Register Variation: A Cross-Linguistic Comparison*, Cambridge: Cambridge University Press. Available at: `tinyurl.com/2qe6fhb5`. (Accessed: 10 July 2022).

Biber, D., Conrad, S., Finegan, E., Johansson, S., Leech, G., Conrad, S., Finegan, E., Quirk, R., Johansson, S., Leech, G., Conrad, S., Finegan, E. and Quirk, R. (2007), *Longman grammar of spoken and written English*, Vol. 2, 6 ed., Edinburgh Gale: Pcarson Education Limited.

Bittner, H. (2020), *Evaluating the evaluator: A novel perspective on translation quality assessment*, New York/London: Routledge, doi: 10.4324/9780367815882.

Bizzoni, Y. and Lapshinova-Koltunski, E. (2021), 'Measuring translationese across Levels of expertise: Are professionals more surprising than students?', *Proceedings of the 23rd Nordic Conference on Computational Linguistics (NoDaLiDa)*, 31 May–2 June, Reykjavik: Linköping University Electronic Press, pp. 53–63. Available at: `aclanthology.org/2021.nodalida-main.6/`. (Accessed: 17 June 2022).

Bjerva, J., Han Veiga, M., Tiedemann, O. and Augenstein, I. (2019), 'What do language representations really represent?', *Computational Linguistics* 45(2), pp. 381–389, doi: 10.1162/COLI.

Blum-Kulka, S. (1986), 'Shifts of cohesion and coherence in translation', in *Interlingual and Intercultural Communication. Discourse and Cognition in Translation and Second Language Acquisition Studies*, Tüübingen: Gunter Narr, pp. 17–35.

Borisova, E., Grinshtein, A. and Protchenko, A. (2019), *A Practical Course of Translation*, 2 ed., Samara: Volga State Academy of Social and Human Sciences. Available at: http://gen.lib.rus.ec/book/index.php?md5=F4407658B903FF70D29524110574DF56

Bouma, G. (2009), 'Normalized (pointwise) mutual information in collocation extraction', *Proceedings of the German Society for Computational Linguistics and Language Technology (GSCL)*, Potsdam, 30 September, pp. 31–40. Available at: svn.spraakdata.gu.se/repos/gerlof/pub/www/Docs/npmi-pfd.pdf. (Accessed: 29 June 2022).

Bowker, L. (2001), 'Towards a methodology for a corpus-based approach to translation evaluation', *Meta: Journal des traducteurs* 46(2), pp. 345–364, doi: 10.7202/002135ar.

Bowker, L. and Bennison, P. (2003), 'Student translation archive and student translation tracking system. Design, development and application', in Zanettin, F., Bernardini, S. and Stewart, D. (eds.) *Corpora in Translator Education*, London: Routledge, pp. 103–118, doi: 10.4324/9781315760162.

Braslavski, P. (2010), 'Marrying relevance and genre rankings: An exploratory study', in *Genres on the Web*, Dordrecht: Springer, pp. 191–208, doi: 10.1007/978-90-481-9178-9_9.

Callison-Burch, C., Fordyce, C., Koehn, P., Monz, C. and Schroeder, J. (2007), '(Meta-)evaluation of machine translation', *Proceedings of the 2nd Workshop on Statistical Machine Translation*, Prague, 23 June, Association for Computational Linguistics, pp. 136–158, doi: 10.3115/1626355.1626373.

Cappelle, B. and Loock, R. (2017), 'Typological differences shining through: The case of phrasal verbs in translated English', in De Sutter, G., Lefer, M.-A. and Delaere, I. (eds.) *Empirical Translation Studies. New Theoretical and Methodological Traditions*, Mouton de Gruyter, pp. 235–264, doi: 10.1515/9783110459586-009.

Carl, M. and Buch-Kromann, M. (2010), 'Correlating translation product and translation process data of professional and student translators', *Proceedings of the 14th Annual conference of the European Association for Machine Translation*, Saint-Raphaël, 27–28

May, European Association for Machine Translation. Available at: aclanthology.org/2010.eamt-1.14.pdf. (Accessed: 16 June 2022).

Carter, D. and Inkpen, D. (2012), 'Searching for poor quality machine translated text: Learning the difference between human writing and machine translations', in Kosseim, L. and Inkpen, D. (eds.) *Advances in Artificial Intelligence – 25th Canadian Conference on Artificial Intelligence*, Lecture Notes in Computer Science, vol. 7310, Springer, pp. 49–60, doi: 10.1007/978-3-642-30353-1_5.

Castagnoli, S. (2009), Regularities and variations in learner translations: A corpus-based study of conjunctive explicitation, PhD thesis, University of Pisa.

Castagnoli, S., Ciobanu, D., Kunz, K., Kübler, N. and Volanschi, A. (2011), 'Designing a learner translator corpus for training purposes', *Corpora, language, teaching, and resources: From theory to practice* 12, pp. 221–248. Available at: citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.631.6247&rep=rep1&type=pdf. (Accessed: 30 June 2022).

Chatzikoumi, E. (2020), 'How to evaluate machine translation: A review of automated and human metrics', *Natural Language Engineering* 26, pp. 137–161, doi: 10.1017/S1351324919000469.

Chesterman, A. (1998), 'Causes, translations, effects', *Target. International Journal of Translation Studies* 10(2), pp. 201–230.

Chesterman, A. (2004), 'Hypotheses about translation universals', in Hansen, G., Malmkjær, K. and Gile, D. (eds.) *Claims, Changes and Challenges in Translation Studies*, Amsterdam/Philadelphia:: John Benjamins Publishing Company, pp. 1–14, doi: 10.1075/btl.50.02che.

Chesterman, A. (2005), 'Consilience in translation studies', *Revista Canaria de Estudios Ingleses* 51, pp. 19–32. Available at: riull.ull.es/xmlui/bitstream/handle/915/18840/RCEI_51_(%202005)_02.pdf?sequence=1. (Accessed: 13 July 2022).

Chesterman, A. (2010), 'Why study translation universals?', *Kiasm. Acta Translatologica Helsingiensia* 1, pp. 38–48. Available at: helda.helsinki.fi/bitstream/handle/101

38/24313/ath_vol_1%20kiasm_30122010.pdf?sequence=1#page=40. (Accessed: 17 July 2022).

Chowdhury, K. D., Espãna-Bonet, C. and van Genabith, J. (2021), 'Tracing source language interference in translation with graph-isomorphism measures', *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, Varna (remote), 6–8 September, INCOMA Ltd., pp. 380–390, doi: 10.26615/978-954-452-072-4_043.

Chowdhury, K. D., Espãna-Bonet, C., van Genabith, J. and Genabith, J. (2020), 'Understanding translationese in multi-view embedding spaces', *Proceedings of the 28th International Conference on Computational Linguistics*, 8–13 December, Barcelona (remote): International Committee on Computational Linguistics, pp. 6056–6062, doi: 10.18653/v1/2020.coling-main.532.

Colina, S. (2008), 'To cite this article: Sonia Colina (2008) Translation Quality Evaluation, The Translator', *The Translator* 14, pp. 97–134, doi: 10.1080/13556509.2008.10799251.

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, É., Ott, M., Zettlemoyer, L. and Stoyanov, V. (2020), 'Unsupervised cross-lingual representation learning at scale', *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Online, 2–10 July, pp. 8440–8451. Available at: aclanthology.org/2020.acl-main.747.pdf. (Accessed: 31 July 2022).

Corpas Pastor, G., Mitkov, R., Afzal, N. and Pekar, V. (2008), 'A corpus-based NLP study of convergence and simplification', *Proceedings of the 8th Conference of the Association for Machine Translation in the Americas (AMTA'08)*, Waikiki, 21–25 October, Association for Computational Linguistics, pp. 75–81. Available at: aclanthology.org/2008.amta-papers.5.pdf. (Accessed: 20 June 2022).

Daems, J., Macken, L. and Vandepitte, S. (2013), 'Quality as the sum of its parts: A two-step approach for the identification of translation problems and translation quality assessment for HT and MT+PE', *Proceedings of MT Summit XIV Workshop on Post-Editing Technology and Practice*, Nice, 2 September, Vol. 2, pp. 63–71. Available at: aclanthology.org/2013.mtsummit-wptp.8.pdf. (Accessed: 13 July 2022).

Daems, J., Vandepitte, S., Hartsuiker, R. J. and Macken, L. (2017), 'Translation methods and experience: A comparative analysis of human translation and post-editing with students and professional translators', *Meta* 2(2), pp. 245–270, doi: 10.7202/1041023ar.

Dai, G. and Xiao, R. (2011), '"SL shining through" in translational language: A corpus-based study of Chinese translation of English passives', *Translation Quarterly* 62, pp. 85–108. Available at: citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.730.4237&rep=rep1&type=pdf. (Accessed: 17 June 2022).

De Sutter, G., Cappelle, B., De Clercq, O., Loock, R. and Plevoets, K. (2017), 'Towards a corpus-based, statistical approach to translation quality: Measuring and visualizing linguistic deviance in student translations', *Linguistica Antverpiensia* 16, pp. 25–39, doi: 10.52034/lanstts.v16i0.440.

Delaere, I. (2015), Do translations walk the line? Visually exploring translated and non-translated texts in search of norm conformity, PhD thesis, Ghent University. Available at: biblio.ugent.be/publication/5888594/file/5888595

Delaere, I., De Sutter, G. and Plevoets, K. (2012), 'Is translated language more standardized than non-translated language? Using profile-based correspondence analysis for measuring linguistic distances between language varieties', *Target* 24(2), pp. 203–224, doi: 10.1075/target.24.2.01del.

Demšar, J. (2006), 'Statistical comparisons of classifiers over multiple data sets', *The Journal of Machine learning research* 7, pp. 1–30.

Denturck, K. (2009), 'Translation universals: the case of causal connectives in French and Dutch translations a corpus-based study', *Proceedings of the Workshop on Connectives across Languages: Explicitation and Grammaticalization of Contingency Relations*, Lisbon, 9–12 September, University of Lisbon.

Diwersy, S., Evert, S. and Neumann, S. (2014), 'A semi-supervised multivariate approach to the study of language variation', in Szmrecsanyi, B. and Wälchli, B. (eds.) *Linguistic variation in text and speech, within and across languages*, Berlin: Mouton de Gruyter, pp. 174–204. Available at: www.stephanie-evert.de/PUB/DiwersyEvertNeumann2012.pdf

Doyle, M. (2003), 'Translation pedagogy and assessment: Adopting ATA's framework for standard error marking', *The ATA Chronicle* 32(11), pp. 21–29. Available at: kavgic.com/srednjoengleski.info/docs/ATA_Standard_Error_Marking_Scheme.pdf. (Accessed: 14 July 2022).

DQF-MQM. (2015), *Multidimensional Quality Metrics (MQM) Definition*, www.qt21.eu/mqm-definition/definition-2015-12-30.html.

Droganova, K., Lyashevskaya, O. and Zeman, D. (2018), 'Data conversion and consistency of monolingual corpora: Russian UD treebanks', *Proceedings of the 17th International Workshop on Treebanks and Linguistic Theories*, Oslo, 13 December, Vol. 155, pp. 53–66. Available at: ufal.mff.cuni.cz/biblio/attachments/2018-droganova-p7646871821297838363.pdf. (Accessed: 12 July 2022).

Duff, A. (1981), *The third language: Recurrent problems of translation into English*, Oxford: Pergamon.

Dugast, C., Turchi, M., Specia, L., Harris, K., Samiotou, A., Niehues, J. and Busemann Stuhlsatzenhausweg, S. (2016), Deliverable number D5.7 Deliverable title Data Management Plan Type Report Status and version, Technical report, QT Launchpad Project. Available at: www.qt21.eu

Eetemadi, S. and Toutanova, K. (2015), 'Detecting translation direction: A cross-domain study', *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Student Research Workshop*, Denver, 31 May–5 June, pp. 103–109, doi: 10.3115/v1/N15-2014.

Esperança-Rodier, E., Brunet-Manquat, F. and Eady, S. (2019), 'ACCOLÉ: A collaborative platform of error annotation for aligned corpora', *Translating and the Computer 41*, London, 22 November, pp. 27–36. Available at: hal.archives-ouvertes.fr/hal-02363208. (Accessed: 13 July 2022).

Espunya, A. (2014), 'The UPF learner translation corpus as a resource for translator training', *Language resources and evaluation* 48(1), pp. 33–43, doi: 10.1007/s10579-013-9260-1.

Evert, S. (2009), 'Corpora and collocations', in Lüdeling, A. and Kyto, M. (eds.) *Corpus linguistics. An international handbook*, Berlin: Mouton de Gruyter, pp. 1211–1248. Available at: stephanie-evert.de/PUB/Evert2007HSK_extended_manuscript.pdf

Evert, S. and Neumann, S. (2017), 'The impact of translation direction on characteristics of translated texts: A multivariate analysis for English and German', *Empirical Translation Studies: New Methodological and Theoretical Traditions* 300, pp. 47–80, doi: 10.1515/9783110459586-003.

Eyckmans, J., Anckaert, P. and Segers, W. (2009), 'The perks of norm-referenced translation evaluation', in Angelelli, C. V. and Jacobson, H. E. (eds.) *Testing and assessment in translation and interpreting studies*, Amsterdam / Philadelphia: John Benjamins Publishing Company, pp. 73–93, doi: 10.1075/ata.xiv.06eyc.

Fabricius-Hansen, C. (1999), 'Information packaging and translation: Aspects of translational sentence splitting (German–English/Norwegian)', in Doherty, M. (ed.) *Sprachspezifische Aspekte der Informationsverteilung*, Studia grammatica, Akademie Verlag, pp. 175–214, doi: 10.1515/9783050078137-008.

Fawcett, P. (1981), 'Teaching translation theory', *Meta: Journal des traducteurs* 26(2), pp. 141, doi: 10.7202/002734ar.

Federmann, C. (2010), 'Appraise: An open-source toolkit for manual phrase-based evaluation of translations', *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, 17–23 May, pp. 1731–1734. Available at: www.lrec-conf.org/proceedings/lrec2010/pdf/197_Paper.pdf. (Accessed: 31 July 2022).

Fictumova, J. (2017), 'Teaching specialized translation. Error-tagged translation learner corpora', *Sendebar* 1(28), pp. 209–241. Available at: revistaseug.ugr.es/index.php/sendebar/article/download/5419/5647. (Accessed: 09 July 2022).

Firbas, J. (1992), *Functional sentence perspective in written and spoken communication*, Cambridge University Press.

Fomicheva, M., Sun, S., Fonseca, E., Zerva, C., Blain, F., Chaudhary, V., Guzmán, F., Lopatina, N., Specia, L. and Martins, A. F. (2020), 'MLQE-PE: A multilingual quality estimation and post-editing dataset', *arXiv preprint arXiv:2010.04480* .

Fomicheva, M., Sun, S., Yankovskaya, L., Blain, F., Guzmán, F., Fishel, M., Aletras, N., Chaudhary, V. and Specia, L. (2020), 'Unsupervised quality estimation for neural machine translation', *Transactions of the Association for Computational Linguistics* 8, pp. 539–555, doi: 10.1162/tacl_a_00330.

Fonseca, E., Yankovskaya, L., Martins, A. F. T., Fishel, M. and Federmann, C. (2019), 'Findings of the WMT 2019 shared tasks on quality estimation', *Proceedings of the Fourth Conference on Machine Translation. Volume 3: Shared Task Papers*, Florence, 1–2 August, pp. 1–10. Available at: aclanthology.org/W19-5401v2.pdf. (Accessed: 17 July 2022).

Fraser, B. (2006), 'Towards a theory of discourse markers', in Fischer, K. (ed.) *Approaches to discourse particles*, Vol. 1 of *Studies in Pragmatics*, Elsevier Ltd., pp. 189–204, doi: 10.1163/9780080461588_012.

Frawley, W. (1984), 'Prolegomenon to a theory of translation', in Frawley, W. (ed.) *Translation: Literary, Linguistic and Philosophical Perspectives*, Newark: University of Delaware Press, pp. 159–175.

Fung, P. and Cheung, P. (2004), 'Multi-level bootstrapping for extracting parallel sentences from a quasi-comparable corpus', *Proceedings of the 20th international conference on Computational Linguistics*, Geneva, 23–27 August, Association for Computational Linguistics, pp. 1051–1057, doi: 10.3115/1220355.1220506.

Gellerstam, M. (1986), 'Translationese in Swedish novels translated from English', *Translation studies in Scandinavia* pp. 88–95.

Gil, J. R. B. and Pym, A. (2006), 'Technology and translation (a pedagogical overview)', *Translation technology and its teaching* pp. 5–19, doi: 10.1002/9781405198431.wbeal1222. Available at: tinyurl.com/26bgrvhz. (Accessed: 13 July 2022).

Girardi, C., Bentivogli, L., Farajian, M. A. and Federico, M. (2014), 'MT-EquAl: A toolkit for human assessment of machine translation output', *Proceedings of the 25th International Conference on Computational Linguistics: System Demonstrations (COLING 2014)*, Dublin, 23-29 August, pp. 120–123. Available at: aclanthology.org/C14-2026.pdf. (Accessed: 31 July 2022).

Gisev, N., Bell, J. S. and Chen, T. F. (2013), 'Interrater agreement and interrater reliability: Key concepts, approaches, and applications', *Research in Social and Administrative Pharmacy* 9(3), pp. 330–338, doi: 10.1016/j.sapharm.2012.04.004.

Graham, Y. (2015), 'Improving evaluation of machine translation quality estimation', *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, Beijing, 27 July–31 July, pp. 1804–1813, doi: 10.3115/v1/P15-1174.

Graham, Y., Baldwin, T., Moffat, A. and Zobel, J. (2013), 'Continuous measurement scales in human evaluation of machine translation', *Proceedings of the 7th Linguistic Annotation Workshop & Interoperability with Discourse*, Sofia, 11 February, pp. 33–41. Available at: aclanthology.org/W13-2305.pdf. (Accessed: 14 July 2022).

Graham, Y., Baldwin, T., Moffat, A. and Zobel, J. (2015), 'Can machine translation systems be evaluated by the crowd alone', *Natural Language Engineering* 23(1), pp. 3–30, doi: 10.1017/S1351324915000339.

Graham, Y., Haddow, B. and Koehn, P. (2020), 'Statistical power and translationese in machine translation evaluation', in Cohn, T., He, Y. and Liu, Y. (eds.) *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, Online, 16–20 November, Association for Computational Linguistics, pp. 72–81, doi: 10.18653/v1/2020.emnlp-main.6.

Granger, S. and Bestgen, Y. (2017), 'Using collgrams to assess L2 phraseological development: A replication study', in de Haan, P. d. V. and van Vuuren, S. (eds.) *Language, Learners and Levels: Progression and Variation*, Louvain-la-Neuve: Presses

universitaires de Louvain, pp. 385–408. Available at: www.i6doc.com/en/book/?gcoi=28001100826850

Granger, S. and Lefer, M.-A. (2018), 'The Translation-oriented annotation system: A tripartite annotation system for translation research', *International Symposium on Parallel Corpora (ECETT–PaCor)*, Madrid, 5–7 November, pp. 61–63.

Gries, S. (2010), 'Useful statistics for corpus linguistics', in Sánchez, A. and Almela, M. (eds.) *A Mosaic of Corpus Linguistics: Selected Approaches*, Frankfurt am Main: Peter Lang, pp. 269–291. Available at: citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.160.9846{&}rep=rep1{&}type=pdf

Guzmán, F., Chen, P.-J., Ott, M., Pino, J., Lample, G., Koehn, P., Chaudhary Marc', V. and Ranzato, A. (2019), 'The FLORES evaluation datasets for low-resource machine translation: Nepali-English and Sinhala-English', *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, Hong Kong, 3–7 November, doi: 10.18653/v1/D19-1632.

Halliday, M. and Hasan, R. (1989), *Language, context, and text: aspects of language in a social-semiotic perspective*, 2 ed., Oxford: Oxford University Press.

Hansen-Schirra, S. (2011), 'Between normalization and shining-through. Specific properties of English-German translations and their influence on the target language', in Kranich, S., Becher, V., Höder, S. and House, J. (eds.) *Multilingual discourse production: Diachronic and synchronic perspectives*, Vol. 12, Amsterdam: John Benjamins Publishing Company, pp. 133–162, doi: 10.1075/hsm.12.07han.

Hansen-Schirra, S., Steiner, E. and Neumann, S. (2012), *Insights from the Language Pair English-German. Cross-linguistic corpora for the study of translations*, Berlin/Boston: Walter de Gruyter GmbH.

Hassan, H., Aue, A., Chen, C., Chowdhary, V., Clark, J., Federmann, C., Huang, X., Junczys-Dowmunt, M., Lewis, W., Li, M., Liu, S., Liu, T.-Y., Luo, R., Menezes, A., Qin, T., Seide, F., Tan, X., Tian, F., Wu, L., Wu, S., Xia, Y., Zhang, D., Zhang, Z. and Zhou, M. (2018), Achieving human parity on automatic Chinese to English news

translation, Technical report, Microsoft AI & Research. Available at: www.microsoft.com/en-us/research/uploads/prod/2018/03/final-achieving-human.pdf

He, P., Gao, J. and Chen, W. (2021), 'DeBERTaV3: Improving DeBERTa using ELECTRA-style pre-training with gradient-disentangled embedding sharing'. Available at: https://arxiv.org/abs/2111.09543

Heafield, K. (2011), 'KenLM: faster and smaller language model queries', in Callison-Burch, C., Koehn, P., Monz, C. and Zaidan, O. F. (eds.) *Proceedings of the EMNLP 2011 Sixth Workshop on Statistical Machine Translation*, Association for Computational Linguistics, pp. 187–197. Available at: aclanthology.org/W11-2123.pdf. (Accessed: 28 June 2022).

Hegrenæs, C. F. and Simonnæs, I. (2020), 'Using rubrics for translation assessment: The case of the national translator accreditation exam in Norway', *Translatologia* 1(1), pp. 27–51. Available at: www.translatologia.ukf.sk/2020/02/using-rubrics-for-translation-assessment-the-case-of-the-national-translator-accreditation-exam-in-norway/. (Accessed: 14 July 2022).

House, J. (2001), 'Translation quality assessment: Linguistic description versus social evaluation', *Meta: Journal des traducteurs* 46(2), pp. 243, doi: 10.7202/003141ar.

Hu, H. and Kübler, S. (2021), 'Investigating Translated Chinese and Its Variants Using Machine Learning', *Natural Language Engineering* 27(3), pp. 339–372, doi: 10.1017/S1351324920000182.

Hu, H., Li, W. and Kübler, S. (2018), 'Detecting syntactic features of translated Chinese', *Proceedings of the 2nd Workshop on Stylistic Variation*, New Orleans, 1–6 June, Association for Computational Linguistics, pp. 20–28, doi: 10.18653/v1/W18-1603.

Huertas-Barros, E., Vandepitte, S. and Iglesias-Fernandez, E. (2018), *Quality assurance and assessment practices in translation and interpreting*, Advances in Linguistics and Communication Studies, Hershey: IGI Global, doi: 10.4018/978-1-5225-5225-3.

Ilisei, I. (2012), A machine learning approach to the identification of translational language: an inquiry into translationese, PhD thesis, University of Wolverhampton. Available at: clg.wlv.ac.uk/papers/ilisei-thesis.pdf

Ilisei, I., Inkpen, D., Pastor, G. C. and Mitkov, R. (2010), 'Identification of translationese: A machine learning approach', *International Conference on Intelligent Text Processing and Computational Linguistics*, Iasi, 21–27 March, Berlin/Heidelberg: Springer, pp. 503–511, doi: 10.1007/978-3-642-12116-6_43.

Ive, J., Blain, F. and Specia, L. (2018), 'deepQuest: A framework for neural-based quality estimation', *Proceedings of the 27th International Conference on Computational Linguistics*, pp. 3146–3157. Available at: aclanthology.org/C18-1266.pdf. (Accessed: 14 July 2022).

Jiménez-Crespo, M. A. (2009), 'The evaluation of pragmatic and functionalist aspects in localization: Towards a holistic approach to quality assurance', *The Journal of Internationalization and Localization* 1(1), pp. 60–93, doi: 10.1075/jial.1.03jim.

Jing, Y. and Liu, H. (2015), 'Mean hierarchical distance augmenting mean dependency distance', in Nivre, J. and Hajicova, E. (eds.) *Proceedings of the Third International Conference on Dependency Linguistics (Depling 2015)*, Uppsala, 24–26 August, pp. 161–170. Available at: aclanthology.org/W15-2119.pdf. (Accessed: 26 June 2022).

Johansson, S. (2008), 'Contrastive analysis and learner language: A corpus-based approach'. Available at: www.hf.uio.no/ilos/english/research/groups/English_Language_and_Corpus_Linguistics_Research/papers/contrastive-analysis-and-learner-language_learner-language-part.pdf

Karakanta, A. and Teich, E. (2019), 'Detecting and analysing translationese with probabilistic language models', (Conference presentation).

Karakanta, A., Vela, M. and Teich, E. (2018), 'Europarl-UDs: Preserving and extending metadata in parliamentary debates', *Proceedings of the LREC 2018 Workshop "ParlaCLARIN: Creating and Using Parliamentary Corpora"*, Miyazaki, 7 May, pp. 33–38. Available at: lrec-conf.org/workshops/lrec2018/W2/pdf/book_of_proceedings.pdf#page=44. (Accessed: 09 July 2022).

Katinskaya, A. and Sharoff, S. (2015), 'Applying multi-dimensional analysis to a Russian webcorpus: Searching for evidence of genres', in Piskorski, J., Pivovarova, L., Šnajder,

J., Tanev, H. and Yangarber, R. (eds.) *Proceedings of the 5th Workshop on Balto-Slavic Natural Language Processing*, Hissar, 5-11 September, pp. 65–74. Available at: www.aclweb.org/anthology/W15-5311. (Accessed: 25 June 2022).

Kepler, F., Trénous, J., Treviso, M., Vera, M., André, U. and Martins, F. T. (2019), 'OpenKiwi: An open source framework for quality estimation', *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, Florence, 28 July–2 August, pp. 117–122. Available at: aclanthology.org/P19-3.pdf#page=131. (Accessed: 14 July 2022).

Kilgarriff, A., Avinesh, P. V. S. and Pomikálek, J. (2011), 'Comparable corpora BootCaT', *Electronic lexicography in the 21st century: New applications for new users. Proceedings of eLex 2011*, Bled, 10–12 November. Available at: www.sketchengine.eu/wp-content/uploads/Comparable_Corpora_BootCaT_2011.pdf. (Accessed: 16 June 2022).

Kim, H., Jung, H. Y., Kwon, H., Lee, J. H. and Na, S. H. (2017), 'Predictor-estimator: Neural quality estimation based on target word prediction for machine translation', *ACM Transactions on Asian and Low-Resource Language Information Processing* 17(1), pp. 1–22, doi: 10.1145/3109480.

Kim, H., Lee, J.-H. and Na, S.-H. (2017), 'Predictor-estimator using multilevel task learning with stack propagation for neural quality estimation', *Proceedings of the Second Conference on Machine Translation*, Copenhagen, 7–8 September, pp. 562–568. Available at: aclanthology.org/W17-4763.pdf. (Accessed: 17 July 2022).

Klubička, F., Toral, A. and Sánchez-Cartagena, V. M. (2018), 'Quantitative fine-grained human evaluation of machine translation systems: A case study on English to Croatian', *Machine Translation* 32(3), pp. 195–215. Available at: www.jstor.org/stable/pdf/44987869.pdf. (Accessed: 31 July 2022).

Koby, G. S. and Baer, B. J. (2005), 'From professional certification to the translator training classroom: Adapting the ATA error marking scale', *Translation Watch Quarterly* 1, pp. 33–45. Available at: www.translocutions.com/tsi/twq/translation_watch_quarterly_December2005_issue1_sample.pdf#page=33. (Accessed: 14 July 2022).

Koehn, P. et al. (2005), 'Europarl: A parallel corpus for statistical machine translation', *Machine Translation Summit*, Phuket, 12–16 September, Association for Computational Linguistics, pp. 79–86. Available at: aclanthology.org/2005.mtsummit-papers.11.pdf. (Accessed: 17 June 2022).

Koponen, M. (2010), 'Assessing machine translation quality with error analysis', *Electronic proceedings of the VIII KäTu symposium on translation and interpreting studies*, Vol. 4, pp. 1–12. Available at: www.sktl.fi/@Bin/40701/Koponen_MikaEL2010.pdf. (Accessed: 13 July 2022).

Koppel, M. and Ordan, N. (2011), 'Translationese and its dialects', *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, Vol.1*, Association for Computational Linguistics, pp. 1318–1326. Available at: aclanthology.org/P11-1132.pdf. (Accessed: 23 June 2022).

Korzen, I. and Gylling, M. (2011), 'What can contrastive linguistics tell us about translating discourse structure?', *Proceedings from Contrastive Linguistics, Translation Studies, Machine Translation – What can we learn from each other? (GSCL2011)*, 27 September, Hamburg, pp. 5–11. Available at: core.ac.uk/download/pdf/17278774.pdf. (Accessed: 26 June 2022).

Kozlova, A., Shmatova, M. and Frolov, A. (2016), 'YSDA participation in the WMT'16 quality estimation shared task', *Proceedings of the First Conference on Machine Translation*, Berlin, 11–12 August, Vol. 2, pp. 793–799. Available at: aclanthology.org/W16-2385.pdf. (Accessed: 17 July 2022).

Kruger, H. and Rooy, B. (2012), 'Register and the features of translated language', *Across Languages and Cultures* 13(1), pp. 33–65, doi: 10.1556/Acr.13.2012.1.3.

Kruger, H. and van Rooy, B. (2010), 'The features of non-literary translated language: A pilot study', *Proceedings of the International Symposium on Using Corpora in Contrastive and Translation Studies*, Ormskirk, 27–29 July, pp. 1–29. Available at: www.lancaster.ac.uk/fass/projects/corpus/UCCTS2010Proceedings/papers/Kruger_Van_%20Rooy.pdf. (Accessed: 23 June 2022).

Kübler, N., Mestivier, A. and Pecman, M. (2018), 'Teaching specialised translation through corpus linguistics: Translation quality assessment and methodology evaluation and enhancement by experimental approach', *Meta: Journal des traducteurs* 63(3), pp. 807–825. Available at: www.erudit.org/en/journals/meta/1900-v1-n1-meta04634/1060174ar.pdf. (Accessed: 13 July 2022).

Kunilovskaya, M. (2015), 'How far do we agree on the quality of translation?', *English Studies at NBU* 1(1), pp. 18–31, doi: 10.33919/esnbu.15.1.2.

Kunilovskaya, M. (2017), 'Linguistic tendencies in English to Russian translation: The case of connectives', *Computational Linguistics and Intellectual Technologies: papers from the annual conference "Dialogue"*, Moscow, 31 May–3 June, Vol. 2, pp. 221–233. Available at: ://www.dialog-21.ru/media/3920/kunilovskayama.pdf. (Accessed: 16 April 2021).

Kunilovskaya, M., Ilyushchenya, T., Morgoun, N. and Mitkov, R. (2022), 'Source language difficulties in learner translation: Evidence from an error-annotated corpus', *Target* , doi: 10.1075/target.20189.kun.

Kunilovskaya, M. and Kutuzov, A. (2018), 'Universal Dependencies-based syntactic features in detecting human translation varieties', in Hajič, J. (ed.) *Proceedings of the 16th International Workshop on Treebanks and Linguistic Theories*, Prague, 23–24 January, Association for Computational Linguistics, pp. 27–36. Available at: aclanthology.org/W17-7606.pdf. (Accessed: 12 July 2022).

Kunilovskaya, M. and Lapshinova-Koltunski, E. (2019), 'Translationese features as indicators of quality in English-Russian human translation', *Proceedings of the Human-Informed Translation and Interpreting Technology Workshop (HiT-IT 2019)*, Varna, 5-6 September, INCOMA Ltd., pp. 47–56, doi: 10.26615/issn.2683-0078.2019_006.

Kunilovskaya, M. and Lapshinova-Koltunski, E. (2020), 'Lexicogrammatic translationese across two targets and competence levels', *Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020)*, Marseille (online), 13–15 June, The European Language Resources Association (ELRA), pp. 4102–4112. Available at: aclanthology.org/2020.lrec-1.505. (Accessed: 16 April 2021).

Kunilovskaya, M., Lapshinova-Koltunski, E. and Mitkov, R. (2021*a*), 'Fiction in Russian translation: A translationese study', *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, Varna (online), 6–8 September, INCOMA Ltd., pp. 739–747, doi: 10.26615/978-954-452-072-4_084.

Kunilovskaya, M., Lapshinova-koltunski, E. and Mitkov, R. (2021*b*), 'Translationese in Russian literary texts', *Proceedings of the 5th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, Punta Cana (online), 7–11 November, Association for Computational Linguistics, pp. 101–112, doi: 10.18653/v1/2021.latechclfl-1.12.

Kunilovskaya, M., Morgoun, N. and Pariy, A. (2018), 'Learner vs. professional translations into Russian: Lexical profiles', *Translation and Interpreting* 10(1), pp. 33–52, doi: 10.12807/ti.110201.2018.a03.

Kunilovskaya, M. and Pastor, G. C. (2021), 'Translationese and register variation in English-to-Russian professional translation', in Wang, V. X., Li, D. and Lim, L. (eds.) *New Frontiers in Translation Studies*, Singapore: Springer Nature, pp. 133–180, doi: 10.1007/978-981-16-4918-9_6.

Kunilovskaya, M. and Sharoff, S. (2019), 'Towards functionally similar corpus resources for translation', *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019)*, Varna, 2–4 September, INCOMA Ltd., pp. 583–592.

Kunilovskaya, M., Taslimipoor, S. and Ilyushchenya, T. (2019), 'Functional text representations for building cross-linguistic comparable corpora in English and Russian', *Proceedings of the 12th Workshop on Building and Using Comparable Corpora (BUCC 2019)*, Varna, 5 September, pp. 37–45.

Kunz, K., Degaetano-Ortlieb, S., Lapshinova-Koltunski, E., Menzel, K. and Steiner, E. (2017), 'English-German contrasts in cohesion and implications for translation', in *Empirical Translation Studies. New Theoretical and Methodological Traditions*, Berlin: Walter de Gruyter, pp. 265–312, doi: 10.1515/9783110459586-010.

Kurokawa, D., Goutte, C. and Isabelle, P. (2009), 'Automatic detection of translated text and its impact on machine translation', *Proceedings of MT-Summit XII*, pp. 81–88. Available at: www.researchgate.net/publication/47683937. (Accessed: 22 June 2022).

Kutuzov, A., Kopotev, M., Sviridenko, T. and Ivanova, L. (2016), 'Clustering comparable corpora of Russian and Ukrainian academic texts: Word embeddings and semantic fingerprints', *Proceedings of the 9th Workshop on Building and Using Comparable Corpora*, Portorož, 23 May, pp. 3–10. Available at: www.lrec-conf.org/proceedings/lrec2016/workshops/LREC2016Workshop-BUCC2016_Proceedings.pdf#page=13. (Accessed: 09 July 2022).

Kutuzov, A. and Kunilovskaya, M. (2014), 'Russian Learner Translator Corpus: Design, research potential and applications', in Sojka, P., Horák, A., Kopeček, I. and Pala, K. (eds.) *Proceedings of the 17th International Conference Text, Speech and Dialogue*, Vol. 8655 of *Lecture Notes in Computer Science, vol. 8655*, Springer, pp. 315–323. Available at: www.tsdconference.org/tsd2014/download/preprints/690.pdf

Lapshinova-Koltunski, E. (2013), 'VARTRA: a comparable corpus for analysis of translation variation', *Proceedings of the 6th Workshop on Building and Using Comparable Corpora*, Sofia, 8 August, pp. 77–86. Available at: aclanthology.org/W13-2510. (Accessed: 30 June 2022).

Lapshinova-Koltunski, E. (2017), 'Exploratory analysis of dimensions influencing variation in translation. The case of text register and translation method', in De Sutter, G., Lefer, M.-A. and Isabelle, D. (eds.) *Empirical Translation Studies. New Theoretical and Methodological Traditions*, Trends in Linguistics. Studies and Monographs, vol. 300, Berlin/Boston: De Gruyter Mouton, pp. 207–234, doi: 10.1515/9783110459586-008.

Lapshinova-Koltunski, E. (2019), 'Exploring linguistic differences between novice and professional translators with text classification', in *New Empirical Perspectives on Translation and Interpreting*, Advances in Translation and Interpreting Studies, New York: Routledge, pp. 215–238, doi: 10.4324/9780429030376-10.

Lapshinova-Koltunski, E. and Vela, M. (2015), 'Measuring 'Registerness' in Human and Machine Translation: A Text Classification Approach', *Proceedings of the Second Workshop on Discourse in Machine Translation*, September, pp. 122–131.

Lapshinova-Koltunski, E. and Zampieri, M. (2018), 'Linguistic features of genre and method variation in translation: A computational perspective', in Legallois, D., Charnois, T. and Larjavaara, M. (eds.) *The Grammar of Genres and Styles: From Discrete to Non-Discrete Units*, Vol. 320, Berlin: Mouton de Gruyter, pp. 92–112, doi: 10.1515/9783110595864-005.

Latyshev, L. and Semenov, A. (2003), *Translation: Theory, Practice and Methodology of Teaching*, Moscow: Academiya.

Läubli, S., Castilho, S., Neubig, G., Sennrich, R., Shen, Q. and Toral, A. (2020), 'A set of recommendations for assessing human-machine parity in language translation', *Journal of Artificial Intelligence Research* 67, pp. 653–672, doi: 10.1613/JAIR.1.11371.

Läubli, S., Sennrich, R. and Volk, M. (2018), 'Has machine translation achieved human parity? A case for document-level evaluation', *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, Brussels, 2–4 November, Association for Computational Linguistics, pp. 4791–4796, doi: 10.18653/v1/D18-1512.

Laviosa, S. (1998), 'Core Patterns of Lexical Use in a Comparable Corpus of English Narrative Prose', *Meta: Journal des traducteurs* 43(4), pp. 557, doi: 10.7202/003425ar.

Laviosa, S. (2008), 'Description in the translation classroom: Universals as a case in point', in Pym, A., Shlesinger, M. and Simeoni, D. (eds.) *Beyond descriptive translation studies: Investigations in homage to Gideon Toury*, Amsterdam/Philadelphia: John Benjamins Publishing Company, pp. 119–132, doi: 10.1075/btl.75.10lav.

Lee, D. (2001), 'Genres, registers, text types, domains, and styles: clarifying the concepts and navigating a path through the BNC jungle', *Language Learning and Technology* 5(3), pp. 37–72. Available at: `ro.uow.edu.au/cgi/viewcontent.cgi?article=1615&context=artspapers`. (Accessed: 16 June 2022).

Lembersky, G., Ordan, N. and Wintner, S. (2012), 'Language models for machine translation: Original vs. translated texts', *Computational Linguistics* 38(4), pp. 799–825, doi: 10.1162/COLI_a_00111.

Lembersky, G., Ordan, N. and Wintner, S. (2013), 'Improving statistical machine translation by adapting translation models to translationese', *Association for Computational Linguistics* 39(4), doi: 10.1162/COLI_a_00159.

Li, B., Gaussier, E. and Yang, D. (2018), 'Measuring bilingual corpus comparability', *Natural Language Engineering* 24(4), pp. 523–549, doi: 10.1017/S1351324917000481.

Lijffijt, J. and Nevalainen, T. (2017), 'A simple model for recognizing core genres in the BNC', in Hiltunen, T., McVeigh, J. and Säily, T. (eds.) *Big and Rich Data in English Corpus Linguistics: Methods and Explorations*, Vol. 19 of *Studies in Variation, Contacts and Change in English*, University of Helsinki. Available at: varieng.helsinki.fi/series/volumes/19/lijffijt_nevalainen/

LISA. (2004), *Localization Industry Primer*, 2 ed. Available at: www.immagic.com/eLibrary/ARCHIVES/GENERAL/LISA/L030625P.pdf

Liu, D. (2008), 'Linking adverbials: An across-register corpus study and its implications', *International Journal of Corpus Linguistics* 13(4), pp. 491–518, doi: 10.1075/ijcl.13.4.05liu.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L. and Stoyanov, V. (2019), 'RoBERTa: A Robustly Optimized BERT pretraining approach'. Available at: www.cs.princeton.edu/~danqic/papers/roberta_paper.pdf

Llitjós, A. F., Carbonell, J. G. and Lavie, A. (2005), 'A framework for interactive and automatic refinement of transfer-based machine translation', *Proceedings of the 10th EAMT Conference: Practical applications of machine translation*, Budapest, 30–31 May, pp. 87–96. Available at: aclanthology.org/2005.eamt-1.13.pdf. (Accessed: 31 July 2022).

Lommel, A., Popović, M. and Burchardt, A. (2014), 'Assessing inter-annotator agreement for translation error annotation', *MTE: Workshop on Automatic and Manual Metrics*

*for Operational Translation Evaluation*, Reykjavik, 26 May, Language Resources and Evaluation Conference, pp. 31–37. Available at: www.researchgate.net/publication/277952483. (Accessed: 31 July 2022).

Loock, R. (2016), *La traductologie de corpus*, Villeneuve d'Ascq: Presses universitaires du Septentrion.

Loock, R., Mariaule, M. and Oster, C. (2013), 'Traductologie de corpus et qualité: étude de cas', *Actes du colloque Tralogy II, session 5 – Assessing Quality in MT/Mesure de la qualité en TA*, Paris, 18 January, Tralogy. Available at: hal.archives-ouvertes.fr/hal-02497533/document. (Accessed: 16 June 2022).

Martins, A. F., Junczys-Dowmunt, M., Kepler, F. N., Astudillo, R., Hokamp, C. and Grundkiewicz, R. (2017), 'Pushing the limits of translation quality estimation', *Transactions of the Association for Computational Linguistics* 5, pp. 205–218, doi: 10.1162/tacl_a_00056.

Massardo, I., Van Der Meer, J., O'brien, S., Hollowood, F., Aranberri, N. and Drescher, K. (2016), 'MT post-editing guidelines'. Available at: info.taus.net/mt-post-editing-guidelines

Massey, G. and Wieder, R. (2019), 'Quality assurance in translation and corporate communications', in Huertas-Barros, E., Vandepitte, S. and Iglesias-Fernandez, E. (eds.) *Quality assurance and assessment practices in translation and interpreting*, Advances in linguistics and communication studies, Hershey PA: Information Science Reference, pp. 57–87, doi: 10.4018/978-1-5225-5225-3.ch003.

Mathet, Y., Widï, A. and Métivier, J.-P. (2015), 'The Unified and Holistic Method Gamma ($\gamma$) for Inter-Annotator Agreement Measure and Alignment', *Computational linguistics* 41(3), pp. 437–479, doi: 10.1162/COLI.

Mauranen, A. (2004), 'Corpora, universals and interference', in Mauranen, A. and Kujamäki, P. (eds.) *Translation universals: Do they exist*, Amsterdam: John Benjamins Publishing Company, pp. 65–82, doi: 10.1075/btl.48.07mau.

McEnery, A. and Xiao, Z. (2007), 'Parallel and comparable corpora: What are they up to?', in Anderman, G. and Rogers, M. (eds.) *Incorporating Corpora. The Linguist*

*and the Translator*, Clevedon, Buffalo, Toronto: Multilingual Matters LTD, pp. 18–31. Available at: eprints.lancs.ac.uk/id/eprint/59/1/corpora_and_translation.pdf

Minacori Vibert, P. (2010), 'Translation assessment: The creation of a computer interface', *Reconceptualizing LSP. Online proceedings of the XVII European LSP Symposium 2009*, Aarhus, 17–21 August, Aarhus School of Business, Aarhus University, pp. 1–21. Available at: www.asb.dk/fileadmin/www.asb.dk/isek/vibert.pdf. (Accessed: 16 June 2022).

Nakamura, S. (2007), 'Comparison of features of texts translated by professional and learner translators', *Proceedings of the 4th Corpus Linguistics conference*, University of Birmingham. Available at: www.birmingham.ac.uk/documents/college-artslaw/corpus/conference-archives/2007/35Paper.pdf. (Accessed: 23 June 2022).

Neumann, S. (2013), *Contrastive register variation. A quantitative approach to the comparison of English and German*, Berlin, Boston: Walter de Gruyter.

Nida, E. A. (1964), *Toward a science of translation*, Leiden: Brill.

Nikolaev, D., Karidi, T., Kenneth, N., Mitnik, V., Saeboe, L. and Abend, O. (2020), 'Morphosyntactic predictability of translationese', *Linguistics Vanguard* 6(1), pp. 1–12, doi: 10.1515/lingvan-2019-0077.

Nini, A. (2015), 'Multidimensional Analysis Tagger'.

Nisioi, S. and Dinu, L. P. (2013), 'A clustering approach for translationese identification', *Proceedings of the International Conference Recent Advances in Natural Language Processing (RANLP 2013)*, Hissar, 7–13 September, pp. 532–538. Available at: www.aclweb.org/anthology/R13-1070. (Accessed: 23 June 2022).

Nord, C. (1997), *Translating as a purposeful activity: Functionalist approaches explained*, 3 ed., Manchester: St Jerome. Available at: tinyurl.com/232uh6hj. (Accessed: 13 July 2022).

Nord, C. (2006), 'Translating as a purposeful activity: A prospective approach', *TEFLIN Journal* 17(2), pp. 131–143. Available at: core.ac.uk/download/pdf/233167343.pdf. (Accessed: 16 June 2022).

Novikova, N. (2008), 'Connectives as cohesive devices in an asyndetic composite sentence [Konnektory kak svjazujushhie sredstva v bessojuznom slozhnom predlozhenii]', *Voronezh State University of Architecture and Civil Engineering Herald* 1(10), pp. 92–100.

Olohan, M. (2001), 'Spelling out the optionals in translation: a corpus study', in Rayson, P., Wilson, A., McEnery, T., Hardie, A. and Khoja, S. (eds.) *Proceedings of the Corpus Linguistics*, Lancaster University, pp. 423–432. Available at: tinyurl.com/24cpk57l

Phelan, M. (2017), 'Analytical assessment of legal translation: A case study using the American Translators Association framework', *The Journal of Specialized Translation* 27, pp. 189–210. Available at: jostrans.org/issue27/art_phelan.pdf. (Accessed: 14 July 2022).

Popescu, M. (2011), 'Studying translationese at the character level', *Proceedings of the International Conference Recent Advances in Natural Language Processing (RANLP 2011)*, Hissar, September, pp. 634–639. Available at: aclweb.org/anthology/R11-1091. (Accessed: 17 June 2022).

Popović, M. (2017), 'Comparing language related issues for NMT and PBMT between German and English', *The Prague Bulletin of Mathematical Linguistics* 108(1), pp. 209–220, doi: 10.2478/pralin-2018-0010.PBML.

Popovic, M. (2020), 'On the differences between human translations', in Martins, A., Moniz, H., Fumega, S., Martins, B., Batista, F., Coheur, L., Parra, C., Trancoso, I., Turchi, M., Bisazza, A., Moorkens, J., Guerberof, A., Nurminen, M., Marg, L. and Forcada, M. L. (eds.) *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, Lisboa, 3–5 November, European Association for Machine Translation, pp. 365–374. Available at: aclanthology.org/2020.eamt-1.39.pdf. (Accessed: 09 July 2022).

Priyatkina, A. F., Starodumova, E. A., Sergeeva, G. N. and etc. (2001), *A Russian dictionary of functional words [Slovar' sluzhebnyh slov russkogo jazyka]*, Vladivostok: Far Eastern Federal University.

Puurtinen, T. (2003), 'Genre-specific features of translationese? Linguistic differences between translated and non-translated Finnish children's literature', *Literary and Linguistic Computing* 18(4), pp. 389–406, doi: 10.1093/llc/18.4.389.

Pylypenko, D., Amponsah-Kaakyire, K., Chowdhury, K. D., van Genabith, J. and España-Bonet, C. (2021), 'Comparing feature-engineering and feature-learning approaches for multilingual translationese classification', *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, Online and Punta Cana, 7–11 November, Association for Computational Linguistics, pp. 8596–8611, doi: 10.18653/v1/2021.emnlp-main.676.

Pym, A. (1992), 'Translation error analysis and the interface with language teaching', *The Teaching of Translation* pp. 279–288. Available at: usuaris.tinet.cat/apym/on-line/training/1992_error.pdf. (Accessed: 17 June 2022).

Pym, A. (2003), 'Redefining translation competence in an electronic age. In defence of a minimalist approach', *Meta: Journal des traducteurs* 48(4), pp. 481–497, doi: 10.7202/008533ar.

Rabadán, R., Labrador, B. and Ramón, N. (2009), 'Corpus-based contrastive analysis and translation universals: A tool for translation quality assessment', *Babel* 55(4), pp. 303–328, doi: 10.1075/babel.55.4.01rab.

Rabinovich, E., Ordan, N. and Wintner, S. (2017), 'Found in translation: reconstructing phylogenetic language trees from translations', *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Vancouver, 30 July–4 August, Association for Computational Linguistics, pp. 530–540, doi: 10.18653/v1/P17-1049.

Rabinovich, E. and Wintner, S. (2015), 'Unsupervised identification of translationese', *Transactions of the Association for Computational Linguistics* 3, pp. 419–432. Available at: aclanthology.org/Q15-1030.pdf. (Accessed: 23 June 2022).

Ranasinghe, T., Orasan, C. and Mitkov, R. (2020), 'TransQuest at WMT2020: Sentence-level direct assessment', *Proceedings of the Fifth Conference on Machine Translation (WMT)*, Online, 19–20 November, pp. 1049–1055. Available at: aclanthology.org/2020.wmt-1.122.pdf. (Accessed: 14 July 2022).

Redelinghuys, K. (2016), 'Levelling-out and register variation in the translations of experienced and inexperienced translators: A corpus-based study', *Stellenbosch Papers in Linguistics* 45(1), pp. 189–220, doi: 10.5774/45-0-198.

Redelinghuys, K. and Kruger, H. (2015), 'Using the features of translated language to investigate translation expertise: A corpus-based study', *International Journal of Corpus Linguistics* 20(3), pp. 293–325, doi: 10.1075/ijcl.20.3.02red.

Řehůřek, R. and Sojka, P. (2010), 'Software framework for topic modelling with large corpora', *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, 22 May, Valletta, Malta: ELRA, pp. 45–50. Available at: is.muni.cz/publication/884893/en. (Accessed: 29 June 2022).

Reimers, N. and Gurevych, I. (2019), 'Sentence-BERT: sentence embeddings using siamese BERT-networks', *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, Hong Kong, China: Association for Computational Linguistics, pp. 3982–3992. Available at: https://github.com/UKPLab/

Reiss, K. and Vermeer, H. J. (1984), *Towards a general theory of translational action: Skopos theory explained*, 3 ed., London/New York: Routledge, translated by Christiane Nord, English reviewed by Marina Dudenhöfer. Available at: tinyurl.com/2cvpc5er. (Accessed: 13 July 2022).

Rubino, R., Lapshinova-Koltunski, E. and Van Genabith, J. (2016), 'Information Density and Quality Estimation Features as Translationese Indicators for Human Translation Classification', *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, San Diego, 12–17June, The Association for Computational Linguistics, pp. 960–970, doi: 10.18653/v1/N16-1110.

Saadany, H. and Orasan, C. (2021), 'BLEU, METEOR, BERTScore: Evaluation of metrics performance in assessing critical translation errors in sentiment-oriented text', *Proceedings of the Translation and Interpreting Technology Online Conference TRITON 2021*, Online, 5–7 July, pp. 48–56, doi: 10.26615/978-954-452-071-7_006.

Sanchez-Torron, M. and Koehn, P. (2016), 'Machine translation quality and post-editor productivity', *Proceedings of AMTA Conference* 1, pp. 16–26. Available at: amtaweb.org/wp-content/uploads/2016/10/AMTA2016_Research_Proceedings_v7.pdf#page=22. (Accessed: 14 July 2022).

Santos, D. (1995), 'On grammatical translationese', *Proceedings of the 10th Nordic Conference on Computational Linguistics*, Helsinki, 29–30 May, University of Helsinki, pp. 59–66. Available at: citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.44.8137&rep=rep1&type=pdf. (Accessed: 17 June 2022).

Scarpa, F. (2006), 'Corpus-based quality-assessment of specialist translation: A study using parallel and comparable corpora in English and Italian', in Gotti, M., Sarcevic, S. and Šarčević, S. (eds.) *Insights into specialized translation*, Bern: Peter Lang, pp. 155–172. Available at: tinyurl.com/ygefcko5

Scarton, C. (2016), Document-level machine translation quality estimation, PhD thesis, University of Sheffield. Available at: etheses.whiterose.ac.uk/16763/{%}0Acore.ac.uk/download/pdf/80689294.pdf

Schäffner, C. and Adab, B. (2001*a*), 'The idea of the hybrid texts and translation: Contact as conflict', *Across Languages and Cultures* 2(2), pp. 167–180, doi: 10.1556/Acr.2.2001.2.1.

Schäffner, C. and Adab, B. (2001*b*), 'The idea of the hybrid texts and translation revisited', *Across Languages and Cultures* 2(2), pp. 277–302, doi: 10.1556/Acr.2.2001.2.11.

Secara, A. (2005), 'Translation evaluation - a state of the art survey', *Proceedings of the eCoLoRe/MeLLANGE Workshop*, Leeds, 21–23 March, pp. 39–44. Available at: corpus.leeds.ac.uk/cts/_publications/leeds-cts-2005-s/leeds-cts-2005-secara.pdf. (Accessed: 14 July 2022).

Shah, K., Avramidis, E., Biçici, E. and Specia, L. (2013), 'QuEst – design, implementation and extensions of a framework for machine translation quality estimation', *The Prague Bulletin of Mathematical Linguistics* 100(1), pp. 19–30, doi: 10.2478/pralin-2013-0008.

Sharoff, S. (2018), 'Functional Text Dimensions for annotation of Web corpora', *Corpora* 13(1), pp. 65–95, doi: 10.3366/cor.2018.0136.

Shvedova, N. J. (1980), *Russian Grammar [Grammatika russkogo jazyka]*, Moscow: Science [Nauka]. Available at: rusgram.narod.ru/. (Accessed: 28 June 2022).

Sitchinava, D. (2019), 'On parallel corpora within the Russian National Corpus: New languages and new challenges [Parallel'nye teksty v sostave Nacional'nogo korpusa russkogo jazyka: novye jazyki i novye zadachi]', *Herald of Vinogradov's Institute of the Russian Language [Trudy Instituta russkogo jazyka im. VV Vinogradova]* 21, pp. 41–61, doi: 10.31912/pvrli-2019.21.2.

Sominsky, I. and Wintner, S. (2019), 'Automatic Detection of Translation Direction', in Mitkov, R. and Angelova, G. (eds.) *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019)*, Varna, 2–4 September, INCOMA Ltd., pp. 1131–1140, doi: 10.26615/978-954-452-056-4_130.

Specia, L., Blain, F., Fomicheva, M., Fonseca, E., Chaudhary, V., Guzmán, F. and Martins, A. F. T. (2020), 'Findings of the WMT 2020 shared task on quality estimation', *Proceedings of the Fifth Conference on Machine Translation (WMT)*, Online, 19–20 November, pp. 743–764. Available at: aclanthology.org/2020.wmt-1.79.pdf. (Accessed: 14 July 2022).

Specia, L., Blain, F., Fomicheva, M., Zerva, C., Li, Z., Chaudhary, V. and Martins, A. F. (2021), 'Findings of the WMT 2021 shared task on quality estimation', *Proceedings of the Sixth Conference on Machine Translation (WMT)*, Punta Cana, 10–11 Novermber, pp. 684–725. Available at: aclanthology.org/2021.wmt-1.71.pdf. (Accessed: 14 July 2022).

Specia, L., Blain, F., Logacheva, V., Astudillo, R. F. and Martins, A. (2018), 'Findings of the WMT 2018 shared task on quality estimation', *Proceedings of the Third Conference*

*on Machine Translation (WMT)*, Brussels, 31 October–1 November, Vol. 2, pp. 689–709, doi: 10.18653/v1/W18-64078.

Specia, L., Paetzold, G. H. and Scarton, C. (2015), 'Multi-level translation quality prediction with QUEST++', in Chen, H.-H. and Markert, K. (eds.) *Proceedings of ACL-IJCNLP 2015 System Demonstrations*, Beijing, 27 July, Association for Computational Linguistics, pp. 115–120, doi: 10.3115/v1/p15-4020.

Specia, L., Scarton, C. and Paetzold, G. H. (2018), *Quality estimation for machine translation*, Vol. 11, San Rafael: Morgan & Claypool Publishers, doi: 10.2200/S00854ED1V01Y201805HLT039.

Specia, L., Scarton, C. and Specia, L. (2016), 'Translation quality estimation tutorial hands-on QuEst++'. Available at: `coling2016.anlp.jp/doc/tutorial/slides/T4/QE_Coling2016_handout.pdf`

Stenetorp, P., Pyysalo, S., Topi, G., Ohta, T., Ananiadou, S. and Tsujii, I. (2012), 'BRAT: A web-based tool for NLP-assisted text annotation', *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, Avignon, 23–27 April, Association for Computational Linguistics, pp. 102–107. Available at: `aclanthology.org/E12-2021.pdf`. (Accessed: 09 July 2022).

Štěpánková, K. (2014), Learner Translation Corpus: CELTraC (Czech-English Learner Translation Corpus), Master's thesis, Department of English and American Studies, Faculty of Arts, Masaryk University.

Straka, M. and Straková, J. (2017), 'Tokenizing, POS tagging, lemmatizing and parsing UD 2.0 with UDPipe', in Zeman, D., Hajic, J., Popel, M., Potthast, M., Straka, M., Ginter, F., Nivre, J. and Petrov, S. (eds.) *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, Vancouver, 3–4 August, Association for Computational Linguistics, pp. 88–99, doi: 10.18653/v1/K17-3009.

Stymne, S. (2011), 'Blast: A tool for error analysis of machine translation output', *Proceedings of the ACL-HLT 2011: System Demonstrations*, Portland, 19–24 June, pp. 56–61. Available at: `aclanthology.org/P11-4010.pdf`. (Accessed: 13 July 2022).

Stymne, S. (2017), 'The Effect of Translationese on Statistical Machine Translation', *Proceedings ofthe 21st Nordic Conference ofComputational Linguistics*, pp. 241–246.

Teich, E. (2003), *Cross-Linguistic Variation in System and Text. A Methodology for the Investigation of Translations and Comparable Texts*, Berlin: Mouton de Gruyter.

Tirkkonen-Condit, S. (1990), 'Professional vs. non-professional translation: A think-aloud protocol study', *Learning, Keeping and Using Language* 2, pp. 381–394, doi: 10.1075/z.lkul2.28tir.

Toury, G. (1980), *In search of a theory of translation*, Porter Institute for Poetics and Semiotics, Tel Aviv University.

Toury, G. (1995), *Descriptive Translation Studies and Beyond*, Amsterdam: John Benjamins Publishing Company.

Turner, B., Lai, M. and Huang, N. (2010), 'Error deduction and descriptors–A comparison of two methods of translation test assessment', *Translation & Interpreting* 2(1), pp. 11–23. Available at: trans-int.org/index.php/transint/article/viewFile/42/66. (Accessed: 14 July 2022).

Ustaszewski, M. (2019), 'Exploring adequacy errors in neural machine translation with the help of cross-language aligned word embeddings', *Proceedings of the 2nd Workshop on Human-Informed Translation and Interpreting Technology (HiT-IT 2019)*, Varna, 5–6 September, pp. 122–128. Available at: aclanthology.org/W19-8715.pdf. (Accessed: 31 July 2022).

Van Brussel, L., Tezcan, A. and Macken, L. (2018), 'A fine-grained error analysis of NMT, SMT and RBMT output for English-to-Dutch', *Proceedings of the 11th International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, 7–12 May, pp. 3799–3804. Available at: biblio.ugent.be/publication/8561558/file/8563576. (Accessed: 13 July 2022).

Van Halteren, H. (2008), 'Source language markers in EUROPARL translations', *Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008)*, Manchester, 18–22 August, Vol. 1, Coling 2008 Organizing Committee, pp. 937–944, doi: 10.3115/1599081.1599199.

Varga, D., Halácsy, P., Kornai, A., Nagy, V., Németh, L. and Trón, V. (2007), 'Parallel corpora for medium density languages', *Amsterdam Studies in the Theory and History of Linguistic Science* 292, pp. 247, doi: 10.1075/cilt.292.32var.

Vela, M., Schumann, A.-C. and Wurm, A. (2014), 'Beyond linguistic equivalence. An empirical study of translation evaluation in a translation learner corpus', *Proceedings of the EACL Workshop on Humans and Computer-assisted Translation*, Gothenburg, 26 April, pp. 47–56. Available at: www.aclweb.org/anthology/W/W14/W14-0308.pdf. (Accessed: 13 July 2022).

Venuti, L. (1995), *The Translator's Invisibility: A History of Translation*, London: Routledge.

Vermeer, H. J. (1989), 'Skopos and commission in translational action', in Venuti, L. (ed.) *The Translation Studies Reader*, 3 ed., London/New York: Routledge, pp. 191–202. Available at: translationjournal.net/images/e-Books/PDF_Files/The%20Translation%20Studies%20Reader.pdf

Verplaetse, H. and Lambrechts, A. (2019), 'Translation error type in medical and legal student translations: Impact of dictionary, CAT tool and corpus use', *Dragoman Journal of Translation Studies* 8(9), pp. 64–76. Available at: lirias.kuleuven.be/retrieve/547218. (Accessed: 09 July 2022).

Vilar, D., Xu, J., Luis Fernando, D. and Ney, H. (2006), 'Error analysis of statistical machine translation output', *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*, Genoa, 22–28 May, pp. 697–702. Available at: www.lrec-conf.org/proceedings/lrec2006/pdf/413_pdf.pdf. (Accessed: 31 July 2022).

Voita, E., Sennrich, R. and Titov, I. (2019), 'When a good translation is wrong in context: Context-aware machine translation improves on deixis, ellipsis, and lexical cohesion', *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, Florence, 28 July–2 August, Association for Computational Linguistics, pp. 1198–1212, doi: 10.18653/v1/P19-1116.

Volansky, V. (2012), On the Features of Translationese, Master thesis, University of Haifa. Available at: cs.haifa.ac.il/~shuly/publications/vered-thesis.pdf

Volansky, V., Ordan, N. and Wintner, S. (2011), 'More Human or more Translated ? Original Texts vs . Human and Machine Translations', *Proceedings of the Israeli Seminar on Computational Linguistics (ISCOL2011)*.

Volansky, V., Ordan, N. and Wintner, S. (2015), 'On the features of translationese', *Digital Scholarship in the Humanities* 30(1), pp. 98–118, doi: 10.1093/llc/fqt031.

Waddington, C. (2001), 'Should translations be assessed holistically or through error analysis?', *Hermes* 26, pp. 15–37, doi: 10.1515/LES.2004.28.

Williams, M. (2001), 'The Application of Argumentation Theory to Translation Quality Assessment', *Meta* 46(2), pp. 326–344.

Williams, M. (2013), 'A holistic-componential model for assessing translation student performance and competency', *Mutatis Mutandis* 6(2), pp. 419–443. Available at: revistas.udea.edu.co/index.php/mutatismutandis/article/download/17214/15459/62618. (Accessed: 14 July 2022).

Wisniewski, G., Kübler, N. and Yvon, F. (2014), 'A corpus of machine translation errors extracted from translation students exercises', *Proceedings of the Ninth International Conference on Language Resources and Evaluation*, Reykjavik, 26–31 May, pp. 3585–3588. Available at: www.lrec-conf.org/proceedings/lrec2014/pdf/1115_Paper.pdf. (Accessed: 30 June 2022).

Wurm, A. (2013), 'Eigennamen und Realia in einem Korpus studentischer Übersetzungen (KOPTE)', *trans-kom* 6(2), pp. 381–419. Available at: www.trans-kom.eu/bd06nr02/trans-kom_06_02_06_Wurm_Eigennamen.20131212.pdf. (Accessed: 30 June 2022).

Wurm, A. (2020), 'Translation quality in an error-annotated translation learner corpus', in Granger, S. and Lefer, M.-A. (eds.) *Translating and Comparing Languages: Corpus-based Insights*, Louvain-la-Neuve: Presses Universitaires de Louvain, pp. 141–162. Available at: tinyurl.com/254dxnzv

Xiao, R. (2010), 'How different is translated Chinese from native Chinese?', *International Journal of Corpus Linguistics* 15(1), pp. 5–35, doi: 10.1075/ijcl.15.1.01xia.

Xiao, Z. and McEnery, A. (2005), 'Two approaches to genre analysis: Three genres in modern American English', *Journal of English Linguistics* 33(1), pp. 62–82, doi: 10.1177/0075424204273957.

Yankovskaya, L., Tättar, A. and Fishel, M. (2019), 'Quality estimation and translation metrics via pre-trained word and sentence embeddings', *Proceedings of the Fourth Conference on Machine Translation (WMT)*, Florence, 1–2 August, Vol. 3, pp. 101–105. Available at: statmt.org/wmt18/

Yuan, Y. (2018), Human Translation Quality Estimation: Feature-based and deep learning-based, PhD thesis, University of Leeds. Available at: etheses.whiteroseac.uk/26290/

Yuan, Y. and Sharoff, S. (2020), 'Sentence level human translation quality estimation with attention-based neural networks', *Proceedings of the 12th Language Resources and Evaluation Conference*, pp. 1858–1865. Available at: www.atanet.org/certification/

Yuan, Y., Sharoff, S. and Babych, B. (2016), 'MoBiL: A hybrid feature set for automatic human translation quality assessment', *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC 2016)*. Available at: github.com/karlmoritz/bicvm

Zanettin, F. (1998), 'Bilingual comparable corpora and the training of translators', *Meta: Journal des traducteurs* 43(4), pp. 616–630, doi: 10.7202/004638ar.

Zanettin, F. (2013), 'Corpus methods for Descriptive Translation Studies', *Procedia – Social and Behavioral Sciences* 95, pp. 20–32, doi: 10.1016/j.sbspro.2013.10.618.

Zaretskaya, A., Vela, M., Corpas, G. and Seghiri, M. (2016), 'Measuring post-editing time and effort for different types of machine translation errors', *New Voices in Translation Studies* 15, pp. 63–92. Available at: core.ac.uk/download/pdf/196185915.pdf. (Accessed: 17 July 2022).

Zhang, M. and Toral, A. (2019), 'The effect of translationese in machine translation test sets', in Bojar, O., Chatterjee, R., Federmann, C., Fishel, M., Graham, Y., Haddow, B., Huck, M., Yepes, A. J., Koehn, P., Martins, A., Monz, C., Negri, M., Névéol, A., Neves, M., Post, M., Turchi, M. and Verspoor, K. (eds.) *Proceedings of the Fourth Conference on Machine Translation (Volume 1: Research Papers)*, Florence, 1–2 August, Association for Computational Linguistics, pp. 73–81, doi: 10.18653/v1/W19-5208.

Zhou, Y. and Bollegala, D. (2019), 'Unsupervised evaluation of human translation quality', *Proceedings of the 11th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management*, Vienna, 17–19 September, SciTePress-Science and Technology Publications, pp. 55–64. Available at: livrepository.liverpool.ac.uk/3097319/1/Unsupervised%20Evalusation%20of%20Human%20Translation%20Quality.pdf. (Accessed: 31 July 2022).

Zrazhevskaya, T. and Belyaeva, L. (1972), *Problems in English-Russian Translation (based on newspaper articles)*, Moscow: "Mezhdunarodnye Otnosheniya". Available at: http://gen.lib.rus.ec/book/index.php?md5=4FB8D1D26768BF5A981C58ADEE0C5F98

# Appendix A UD features by shorthand code name

1. **acl**

   finite and non-finite clausal modifier of noun (adjectival clause):

   extraction is based on UD default annotation

   e.g. the person *showing* (acl) her around; help people do something to *overcome* (acl) it; людей, *следящих* (acl) за политикой

2. **acl:relcl**

   relative clause:

   a traditional type of adjectival clause, annotated separately in English and Russian UD treebanks. It is a finite clause, which does not have the modified noun. Hu and Kübler (2021) found that there were relative clauses in translations regardless of the translation pair in their multilingual translationese study.

3. **addit**

   additive connectives:

   cumulative frequency of the list items normalised to the number of sentences; see description in Appendix B In a similar recent study, (Hu and Kübler, 2021) reported the accuracy of 96.55% in a translationese classification from a number of SL into Chinese on 160 cohesive markers alone.

4. **adp**

   adpositions:

   all lemmas tagged *ADP* from the lists:

   English: *of, in, unlike, for, at, as, to, along, with, after, on, towards, amongst, within, over, during, by, against, about, out, from, without, into, like, up, between, before, down, across, per, off, around, since, onto, through, beyond, under, despite, than, until, because, upon, among, back, behind, past, outside, throughout, inside, via, above, alongside, versus, below, round*

   Russian: *в, по, за, на, от, с, под, у, из-под, из-за, до, к, о, через, из, над, про, после, вроде, перед, между, насчет, около, внутрь, без, кроме, для, со,*

*при, сквозь, вместо, вокруг, мимо, позади, возле, против, согласно, вдоль, во, среди, напротив, благодаря, помимо, ради, поверх, посреди, меж*

5. **advcl**

   adverbial clause:

   an optional complement which modifies a verb or other predicate (e.g. a temporal clause, consequence, conditional clause, purpose clause, etc.), extracted based on the default dependency tag *advcl*

6. **advers**

   adversative (contrastive) connectives:

   cumulative frequency of the list items normalised to the number of sentences; see description in Appendix B

7. **advmod**

   non-clausal adverbial modifier:

   a modifier of a predicate or a modifier word expressed with an adverb or an adverbial phrase, but not an adpositional phrase; extraction is based on the default UD tag *advmod*.

8. **amod**

   non-clausal adjectival modifier:

   it includes attrib, but is more general and extends to case, such as *There is nothing wrong (amod) with it.*

9. **anysome**

   noun substitutes, i.e. pronouns par excellence, of indefinite and total semantic subtypes:

   English: all lemmas tagged as *PRON* from the list *anybody, anyone, anything, everybody, everyone, everything, somebody, someone, something, elsewhere, everywhere, somewhere, anywhere*

   Russian: all lemmas tagged as *PRON* from the list *некто, нечто, нечего*, lemmas ending in *-то/-нибудь/-либо*, except starting with *какой* and all lemmas from the the list: *кто-кто, кого-кого, кому-кому, кем-кем, ком-ком, что-что, чего-чего, чему-чему, чем-чем, куда-куда, где-где*

10. **appos**

    appositional modifier:

    a modifier of a noun, which serves to define, modify, name, or describe that noun, usually punctuated or parenthetical (e.g. *President Clinton; Sam, my brother*)

11. **attrib**

    adjectives and participles functioning as attributes:

    all words tagged as *ADJ* or having a morphological feature *VerbForm=Part* with the *amod* dependency to their head

    e.g. the *rising* sun; the *coloured* face; *fried* green tomatoes

12. **aux**

    auxiliary:

    a function word associated with a verbal predicate that expresses categories such as tense, mood, aspect, voice or evidentiality; extraction is based on the default dependency tag

13. **aux:pass**

    auxiliary verbs in passive forms:

    extraction is based on UD default tag. See related discussion in ft:passives

14. **case**

    case marking:

    a function of a separate syntactic word, dependent on a noun (including prepositions, postpositions, and clitic case markers) used for any case-marking; extracted using default syntactic tag

15. **caus**

    causative connectives:

    cumulative frequency of the list items normalised to the number of sentences; see description in Appendix B

16. **conj**

    conjunct:

relation between two elements connected by a coordinating conjunction

e.g. *We buy apples, pears (conj), oranges (conj) and bananas (conj).*

17. **ccomp**

    clausal complement:

    extracted using default UD dependency tag e.g. *help people to do (ccomp) smth*; не ожидали, что придет (ccomp)

18. **cconj**

    coordinating conjunctions:

    all listed lemmas tagged *CCONJ*

    English: *and, or, both, yet, either, &, nor, plus, neither, ether*

    Russian: *и, а, или, ни, да, причем, либо, зато, иначе, только, ан, и/или, иль*

19. **compar**

    comparative degree of comparison for adjectives and adverbs:

    synthetic forms are extracted based on the morphological attribute *Degree=Comp*, while analytical forms are counted as adjectives and adverbs with a dependent *more* (English) or *более, больший* (Russian)

20. **compound**

    compound:

    one of the three relations for multiword expressions (MWEs)

    e.g. *ice (compound) cream (compound) flavours*, *в двадцать* (compound) *первом веке*

21. **content_dens**

    lexical density: ratio of PoS disambiguated content word types to all tokens. Content words include lemmas in *ADJ, ADV, VERB, NOUN* PoS categories. PoS disambiguation is achieved by taking into account PoS tags assigned to lemmas: *look_VERB vs look_NOUN*

22. **content_TTR**

    type-to-token ratio based on content words: ratio of PoS disambiguated content word types to their tokens. Content words include lemmas in *ADJ, ADV, VERB,*

*NOUN* PoS categories. PoS disambiguation is achieved by taking into account PoS tags assigned to lemmas: *look_ VERB vs look_ NOUN*

23. **copula**

    copula verbs; lemmas of *be*, *быть, это* that have a *cop* relation to their head, excluding constructions with *there* as head for English

24. **determ**

    pronominal determiners:

    English: all lemmas in the function *det* from the list: *this, some, these, that, any, all, every, another, each, those, either, such*

    Russian: all lemmas in the function *det* from the list: *этот, весь, тот, такой, какой, каждый, любой, некоторый, какой-то, один, сей, это, всякий, некий, какой-либо, какой-нибудь, кое-какой*

25. **deverbals**

    deverbal nouns, names of processes, actions, states:

    English: nouns ending in *-ment, -tion/ -ung, -tion*) or derived by conversion. In the first case the output is filtered with an empirically-driven stop list, including fully substantivised nouns (e.g. *government, population, nation, tuition, etc*). In the second case, a positive filter of 225 lemmas was used to count nounal occurrences of lemmas that also prevail as verbs in our corpus. This excludes items such as *design, set, measure, mark, press, stick, cross, trap, handle.*

    Russian: nouns ending in *-тие, -ение, -ание, -ство, -ция, -ота* filtered with a 150-items stop list to exclude fully substantivised nouns such as *собрание, месторождение, министерство, телевидение, творчество, решение*.

26. **discourse**

    discourse element: a dependency relation for interjections and other discourse particles and elements (which are not clearly linked to the structure of the sentence, except in an expressive way)

27. **epist**

    epistemic stance discourse markers:

cumulative frequency of the list items normalised to the number of sentences; see description in Appendix B

28. **finites**

    verbs in finite form:

    extraction is based on UD default morphological attribute *VerbForm=Fin*

29. **fixed**

    fixed multiword expression:

    used to annotate fixed grammaticised expressions that behave like function words or short adverbials (e.g. *because of, as well as*)

30. **flat**

    flat multiword expression:

    used to annotate exocentric semi-fixed MWEs like names (Hillary Rodham Clinton) and dates (24 December)

31. **infs**

    infinitives:

    This feature captures differences in degrees of nominalisation tendencies.

    English: all cases of a verb form tagged *VerbForm=Inf* with a dependent *to* particle and cases of true bare infinitive, excluding after modal verbs and *have to, going to* and modal adjectival predicates, but including cases after *help, make, bid, let, see, hear, watch, dare, feel.*

    Russian: all occurrences of verb forms with the feature *VerbForm=Inf* except after modal predicates and with the dependent *быть* to exclude future forms (e.g. *отношения будут ухудшаться*)

32. **interrog**

    interrogative sentences:

    all sentences ending in *?*

33. **iobj**

    indirect object:

a function of a nominal phrase which is a core argument of the verb but is not its subject or (direct) object

34. **mark**

    marker:

    a function of a word marking a clause as subordinate to another clause

35. **mdd**

    mean dependency distance (MDD):

    also known as comprehension difficulty (listener's perspective) and measured as "the distance between words and their parents, measured in terms of intervening words" (Hudson, 1995, p.16) as quoted by Jing and Liu (2015, p.162)

36. **mhd**

    mean hierarchical distance (MHD):

    also known as production difficulty (speaker's perspective) : average value of all path lengths travelling from the root to all nodes along the dependency edges (Jing and Liu, 2015, p.164)

37. **mpred**

    modal predicates:

    English: all verbs (except *will/shall*) tagged as *MD* in *XPOS* column of the CoNLL-U annotation file, constructions with *have-to-Inf* and all adjectival modal predicates (given a list of 17 predicatives such as *impossible, likely, sure* with a dependent tagged as *AUX*.

    Russian: lemma *мочь*, lemma *следовать* with a dependent infinitive, three modal adverbs *можно, нельзя, надо* and 11 adjectives from the modal predicative list in the short form identified as *Variant=Short* (e.g. *должен, способный, возможный*)

38. **mquantif**

    adverbial quantifiers:

    all lemmas on pre-defined lists tagged *ADV*.

    English: 37 listed items (e.g. *barely, completely, intensely, almost*)

    Russian: 80 items (e.g. *абсолютно, полностью, сплошь, необыкновенно, достаточно, совершенно, невыносимо, примерно*. For Russian we additionally provide for

functionally similar non-adverbial quantifiers such as *еле, очень, вшестеро, невыразимо, излишне, еле-еле, чуть-чуть, едва-едва, только, капельку, чуточку, едва*.

39. **neg**

    negative particles or main sentence negation:

    counts of lemmas in *no, not, neither, nobody, none, nothing, nowhere* and *нет, не*

40. **nmod**

    nominal modifier:

    a dependency relation of a nominal which is dependent on another noun or a noun phrase and functions as an attribute, or genitive complement

41. **nn**

    total nouns:

    frequency of all lemmas tagged as *NOUN*

42. **nnargs**

    ratio of nouns or proper names as verb arguments:

    core verbal arguments represented by nouns or proper names: ratio of nouns and proper names in the functions of *nsubj, obj, iobj, nsubj:pass* to the total count of these functions

43. **nsubj**

    nominal subject:

    a dependency relation of a nominal which is the syntactic subject and the proto-agent of a clause

44. **numcls**

    cumulative frequency of clauses per sentence:

    number of relations from the list *csubj, acl:relcl, advcl, acl, xcomp, parataxis* annotated in one sentence

45. **nummod**

    numeric modifier:

a dependency relation of a numeral which gives a quantitative characteristic of a noun

46. **obj**

    object:

    a dependency relation of a nominal which functions as the second care argument of a verb after subject.

47. **obl**

    oblique nominal:

    a relation for a nominal (noun, pronoun, noun phrase) functioning as a non-core (oblique) argument or adjunct

48. **parataxis**

    asyndeton:

    a relation between juxtaposed elements of a sentence (paranthetical elements, or parts joined with : or ;). Extraction is based on the default UD annotation.

49. **passives**

    passive constructions in the main predicate:

    This is the only morphological form used by Volansky, Ordan and Wintner (2011). They acknowledge that English is more flexible with forming passives that many other languages, including Russian. Passive is selected as a focus for a corpus-based translationese study on English-to-Chinese material in (Dai and Xiao, 2011). Russian is known for a number of lexico-grammatic agentless structures to convey passive relations between a verb and its arguments. We expect that the less used analytical passive marked with aux:pass will be prevalent in translations to the detriment of the TL-specific forms. English: all verbs tagged *Voice=Pass* with a dependent tagged *aux:pass*;

    Russian: two morphological and two semantic forms of the sentence root verb, if they have the following morphological properties and contexts:

    - morphological attributes *Variant=Short/VerbForm=Part/Voice=Pass* (e.g. *политика была направлена*)

- morphological attributes *VerbForm=Fin/Voice=Pass* (e.g. *война велась*)
- morphological attribute *Voice=Mid* (e.g. *стадион строится на новом месте [the stadium builds-itself in a new location]*)
- sentences where the root verb (except *есть, иметь*) is in third person plural form with no dependent noun in the function of *nsubj* after filtering out a number of exceptions to this general rule (e.g. *стадион возводят на новом месте* [the stadium they-build in a new location], *во Владикавказе ему готовят радушную встречу.* [In Vladikavkas for him they-prepare a warm welcome.])

50. **pasttense**

    verbs in the past tense:

    all occurrences of the feature *Tense=Past*

51. **possp**

    possessive pronouns:

    English: all lemmas in *my, your, his, her, its, our, their* tagged *DET*, or *PRON* and *Poss=Yes*;

    Russian: all lemmas in *мой, твой, ваш, его, ее, её, наш, их, ихний, свой* tagged *DET*

52. **ppron**

    personal pronouns:

    tokens tagged PRON, with any value of attribute *Person=* that do not have *Poss=Yes* feature and are on the list

    English: *i, you, he, she, it, we, they, me, him, her, us, them*

    Russian: *я, ты, вы, он, она, оно, мы, они, меня, тебя, его, её, ее, нас, вас, их, неё, нее, него, них, мне, тебе, ей, ему, нам, вам, им, ней, нему, ним, меня, тебя, него, мной, мною, тобой, тобою, Вами, им, ей, ею, нами, вами, ими, ним, нем, нём, ней, нею*

53. **pverbals**

    participles:

    English: all occurrences of *VerbForm=Part* or *VerbForm=Ger* not in attributive

function *amod* and not part of an analytical form.

Russian: all occurrences of *VerbForm=Part* not in the short form and not in the attributive function, without a dependent auxiliary, and *VerbForm=Conv* (language specific verb form, denoting an adverbial participle/converb) without dependent auxiliary (e.g. *Он ушел, хлопнув дверью.* [He left slamming the door.])

54. **sconj**

    subordinating conjunctions:

    English: all lemmas tagged *SCONJ* in *that, if, as, of, while, because, by, for, to, than, whether, in, about, before, after, on, with, from, like, although, though, since, once, so, at, without, until, into, despite, unless, whereas, over, upon, whilst, beyond, towards, toward, but, except, cause, together*

    Russian: all lemmas tagged *SCONJ* in *что, как, если, чтобы, то, когда, чем, хотя, поскольку, пока, тем, ведь, нежели, ибо, пусть, будто, словно, дабы,раз, насколько, тот, коли, коль, хоть, разве, сколь,ежели, покуда, постольку*

55. **sentlength**

    sentence length:

    number of words per sentence averaged over all sentences in the text.

56. **simple**

    simple sentence:

    frequency of sentences where no words have relations: *csubj, acl:relcl, advcl, acl, xcomp, parataxis*

57. **superl**

    superlative degree of comparison for adjective and adverbs:

    synthetic forms are extracted based on the tag *Degree=Sup*, while analytical forms are counted as adjectives and adverbs with a dependent *most* (for English), and *наиболее, самый* and words starting with *наи-* with the exception of a few homonymous adverbs (e.g. *наискосок*)(for Russian)

58. **tempseq**

    temporal and sequential connectives:

cumulative frequency of the list items normalised to the number of sentences; see description in Appendix B

59. **wdlength**

    word length:

    average number of characters per token, excluding punctuation marks

60. **xcomp**

    a predicative or clausal complement without its own subject:

    it is annotated after phasal verbs (e.g. *started to sing*), in case of infinitive constructions (e.g. *asked me to leave*), etc.; extraction is based on UD default annotation

# Appendix B Semantic groups of discourse markers

Pre-defined lists of discourse markers grouped according to the main semantic relation they mark in our corpora, based on manual analysis.

| | English | Russian |
|---|---|---|
| Additive | 52<br>also, such as, for example, not only, for instance, in particular, moreover, in other words, namely | 52<br>также, при этом, например, кроме того, в частности, к тому же, на самом деле, а именно, иными словами, точнее, причем, вдобавок |
| Adversative | 46<br>still, however, rather than, instead, though, on the other hand, in fact, despite | 34<br>однако, хотя, впрочем, правда, несмотря на, в отличие от, вместе с тем, всё-таки, но на самом деле, наоборот, напротив, зато |
| Causative | 42<br>because, so, due to, so that, therefore, as a result, after all, for this reason, consequently | 49<br>потому, поэтому, поскольку, ведь, так, в результате, ради того, чтобы, затем, что, получается, в этом случае, в связи с тем, дабы, тем более что |
| Temporal/sequential | 110<br>while, since, soon, and then, eventually, further, anyway, thus, at the same time, ultimately, meanwhile | 48<br>пока, наконец, затем, в целом, в то время, как, в заключение, в конце концов, во-первых, в то же время |
| Epistemic markers | 64<br>really, at least, perhaps, of course, probably, in any case, for sure, in reality, no doubt, arguably, clearly, indeed, I/we think, I/we am/are (un) convinced/sure | 86<br>конечно, возможно, может быть, действительно, говорят, на мой взгляд, якобы, полагаю, по сути, в любом случае, кажется, бесспорно, пожалуй |

Table B.1: Semantic groups of listed discourse markers, including the number of items in each group by language and a few examples (ordered by frequency in the originally authored subcorpora)

# Appendix C Core MQM taxonomy

DQF-MQM harmonised error typology[1]

1. Accuracy
    - Addition
    - Mistranslation
    - Omission
    - Untranslated
2. Design
3. Fluency
    - Grammar
    - Grammatical register
    - Inconsistency
    - Spelling
    - Typography
    - Unintelligible
4. Locale convention
5. Style
6. Terminology
7. Verity
    - Completeness
    - Legal requirements
    - Locale-specific content

[1] www.qt21.eu/mqm-definition/definition-2015-12-30.html

# Appendix D *RusLTC* error taxonomy

Hierarchical tags with frequencies

| tag | number |
|---|---|
| Content | 136 |
| content_reference | 380 |
| omission | 59 |
| distortion | 785 |
| nonsense | 391 |
| inexact | 340 |
| unclear | 549 |
| content_cohesion | 429 |
| theme-rheme | 225 |
| logic | 493 |
| content_pragmatics | 454 |
| register | 186 |
| use | 230 |
| SUBTOTAL | 4657 |
| Language | 12 |
| language_lexical | 467 |
| choice-of-word | 1879 |
| combinability | 1144 |
| language_morphology | 87 |
| wordform | 226 |
| language_syntax | 591 |
| incomplete | 88 |
| ungrammatical | 333 |
| word_order | 174 |
| preposition | 88 |
| delete | 552 |
| language_spelling | 245 |
| capitals | 137 |
| typo | 122 |
| language_punctuation | 545 |
| SUBTOTAL | 6690 |
| TOTAL | 11347 |

Table D.1: Distribution of error across categories

Error attributes assigned to spans

| attribute | number |
|---|---|
| Severities | |
| critical | 1159 |
| major | 3034 |
| minor | 7154 |
| Technology | |
| background_info | 67 |
| SL | 81 |
| TL | 144 |
| too_literal | 686 |
| too_free | 59 |
| proper_name | 144 |
| inconsistency | 17 |
| Good_job | 713 |
| AnnotatorNotes | 5548 |

Table D.2: Distribution of error attributes

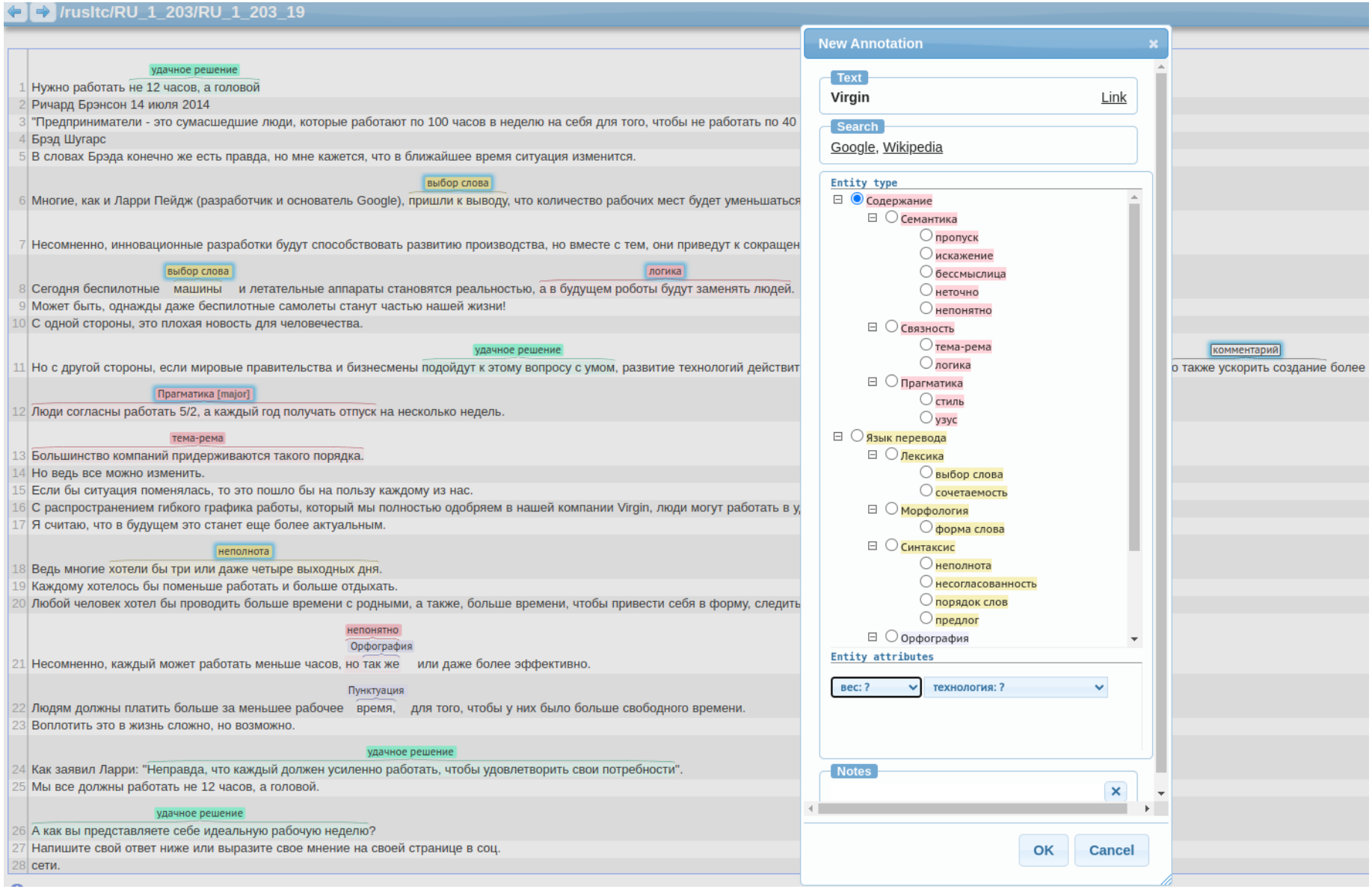



Figure D.1: Screenshot of a *brat*-based annotation environment

| error type | weight |
|---|---|
| Content | 1 |
| content_reference | 2 |
| omission | 2 |
| distortion | 3 |
| nonsense | 3 |
| inexact | 1 |
| unclear | 2 |
| content_cohesion | 2 |
| theme-rheme | 1 |
| logic | 3 |
| content_pragmatics | 1 |
| register | 3 |
| use | 2 |
| Language | 1 |
| language_lexical | 1 |
| choice-of-word | 3 |
| combinability | 3 |
| language_morphology | 1 |
| wordform | 1 |
| language_syntax | 2 |
| incomplete | 3 |
| ungrammatical | 3 |
| word_order | 2 |
| preposition | 3 |
| delete | 2 |
| language_spelling | 1 |
| capitals | 1 |
| typo | 1 |
| language_punctuation | 1 |

Table D.3: Weights assigned to error categories for accuracy and fluency scores

# Appendix E Performance of neural classifiers

| representation | Accuracy | Precision | Recall | F1 |
|---|---|---|---|---|
| chance | 47.72 (+/-5.93) | 47.72 (+/-5.93) | 47.70 (+/-6.00) | 47.58 (+/-5.95) |
| UD | 90.56 (+/-4.14) | 90.59 (+/-4.12) | 90.38 (+/-4.30) | 90.43 (+/-4.22) |
| collgram | 70.82 (+/-5.05) | 70.85 (+/-5.17) | 70.43 (+/-5.11) | 70.33 (+/-5.10) |
| ngram | 69.80 (+/-3.67) | 69.95 (+/-4.45) | 68.99 (+/-3.63) | 69.02 (+/-3.62) |
| **all** | 92.01 (+/-3.78) | 92.04 (+/-3.74) | 91.85 (+/-3.91) | **91.90** (+/-3.84) |
| tfidf | 95.34 (+/-2.04) | 95.31 (+/-2.10) | 95.34 (+/-2.05) | 95.29 (+/-2.05) |
| stsb-xlm-r-m | 93.01 (+/-1.50) | 93.04 (+/-1.53) | 92.91 (+/-1.51) | 92.93 (+/-1.51) |
| TQmono-m | 94.23 (+/-2.76) | 94.31 (+/-2.86) | 94.12 (+/-2.76) | 94.16 (+/-2.78) |
| ruRoberta | 96.34 (+/-2.28) | 96.34 (+/-2.29) | 96.29 (+/-2.30) | 96.30 (+/-2.30) |
| **mdeberta3** | 96.45 (+/-1.84) | 96.47 (+/-1.77) | 96.40 (+/-1.93) | **96.41** (+/-1.87) |

Table E.1: Professional translations: results of translationese classification (neural classifier)

| representation | Accuracy | Precision | Recall | F1 |
|---|---|---|---|---|
| chance | 50.05 (+/-3.72) | 48.65 (+/-3.80) | 48.62 (+/-3.88) | 48.59 (+/-3.84) |
| UD | 90.49 (+/-2.87) | 90.25 (+/-2.95) | 90.17 (+/-3.05) | 90.12 (+/-2.98) |
| collgram | 67.75 (+/-1.62) | 66.63 (+/-2.02) | 64.03 (+/-2.14) | 64.12 (+/-2.37) |
| ngram | 75.81 (+/-5.60) | 76.69 (+/-6.15) | 72.43 (+/-6.55) | 72.91 (+/-6.73) |
| **all** | 92.30 (+/-2.02) | 92.35 (+/-2.03) | 91.76 (+/-2.25) | **91.94** (+/-2.13) |
| tfidf | 94.11 (+/-2.49) | 93.64 (+/-2.57) | 94.68 (+/-2.39) | 93.96 (+/-2.54) |
| stsb-xlm-r-m | 89.54 (+/-3.34) | 89.41 (+/-3.15) | 89.00 (+/-3.66) | 89.06 (+/-3.52) |
| TQmono-m | 95.79 (+/-2.02) | 95.69 (+/-2.05) | 95.65 (+/-2.16) | 95.63 (+/-2.10) |
| ruRoberta | 97.59 (+/-1.78) | 97.50 (+/-1.83) | 97.55 (+/-1.91) | 97.50 (+/-1.86) |
| **mdeberta3** | 97.84 (+/-2.12) | 97.94 (+/-1.92) | 97.61 (+/-2.48) | **97.73** (+/-2.25) |

Table E.2: Student translations: results of translationese classification (neural classifier)

| representation | Accuracy | Precision | Recall | F1 |
|---|---|---|---|---|
| chance | 49.59 (+/-3.47) | 49.45 (+/-3.48) | 49.45 (+/-3.51) | 49.39 (+/-3.49) |
| UD | 78.45 (+/-4.00) | 78.71 (+/-4.10) | 78.24 (+/-3.80) | 78.18 (+/-3.98) |
| collgram | 65.99 (+/-5.39) | 66.42 (+/-6.08) | 64.86 (+/-5.43) | 64.60 (+/-5.58) |
| ngram | 67.34 (+/-4.19) | 67.95 (+/-4.59) | 65.93 (+/-4.54) | 65.57 (+/-5.05) |
| all | 81.02 (+/-2.24) | 81.18 (+/-2.25) | 80.60 (+/-2.18) | 80.71 (+/-2.23) |
| quest61 | 82.66 (+/-4.48) | 83.00 (+/-4.51) | 82.33 (+/-4.47) | 82.38 (+/-4.52) |
| tfidf | 82.51 (+/-5.42) | 82.64 (+/-5.26) | 82.76 (+/-5.29) | 82.45 (+/-5.39) |
| stsb-xlm-r-m | 79.54 (+/-3.13) | 79.65 (+/-3.18) | 79.29 (+/-3.18) | 79.27 (+/-3.21) |
| TQmono-m | 84.55 (+/-4.46) | 84.81 (+/-4.44) | 84.10 (+/-4.60) | 84.26 (+/-4.58) |
| ruRoberta | 89.43 (+/-2.68) | 89.65 (+/-2.70) | 89.18 (+/-2.75) | 89.28 (+/-2.74) |
| **mdeberta3** | 91.06 (+/-3.38) | 91.48 (+/-3.28) | 90.87 (+/-3.39) | **90.93** (+/-3.43) |

Table E.3: Student vs Professional translations (neural classifier)

| representation | Accuracy | Precision | Recall | F1 |
|---|---|---|---|---|
| chance | 51.67 (+/-7.78) | 51.67 (+/-7.80) | 51.67 (+/-7.78) | 51.63 (+/-7.78) |
| **UD** | 56.94 (+/-7.98) | 57.08 (+/-7.95) | 57.01 (+/-7.91) | **56.74** (+/-8.07) |
| collgram | 52.22 (+/-10.45) | 52.45 (+/-10.6) | 52.30 (+/-10.43) | 52.03 (+/-10.4) |
| ngram | 49.17 (+/-9.13) | 49.32 (+/-9.66) | 49.19 (+/-9.33) | 48.66 (+/-9.17) |
| all | 53.89 (+/-8.98) | 53.92 (+/-9.15) | 53.86 (+/-8.98) | 53.60 (+/-9.06) |
| quest61 | 47.22 (+/-8.43) | 47.29 (+/-8.78) | 47.15 (+/-8.48) | 46.88 (+/-8.26) |
| quest17 | 52.50 (+/-8.65) | 52.59 (+/-8.76) | 52.55 (+/-8.68) | 52.19 (+/-8.72) |
| tfidf | 56.39 (+/-6.70) | 56.93 (+/-7.12) | 56.52 (+/-6.77) | 55.76 (+/-6.91) |
| stsb-xlm-r-m | 49.17 (+/-7.86) | 49.09 (+/-8.10) | 49.21 (+/-7.91) | 48.77 (+/-8.09) |
| TQmono-m | 62.50 (+/-4.69) | 63.20 (+/-4.40) | 62.72 (+/-4.35) | 62.11 (+/-5.05) |
| ruRoberta | 72.22 (+/-7.95) | 72.55 (+/-7.79) | 72.33 (+/-7.86) | 72.14 (+/-8.01) |
| **mdeberta3** | 79.44 (+/-6.83) | 79.91 (+/-6.52) | 79.49 (+/-6.84) | **79.32** (+/-6.97) |

Table E.4: Bad vs Good: Binary quality classification (neural classifier)

# Appendix F Comparative-contrastive UD features

Note: In Tables F.1 and F.2 equals sign indicates no statistical differences; * marks cases where there are statistical differences between SL and TL (ref=src)

Table F.1: Properties of UD features sorted by degree of association between values in professional translations and non-translation

| feature | shapiro-tgt | shapiro-ref | bartlett | LogR | trend | means |
|---|---|---|---|---|---|---|
| acl | – | – | – | 0.62 | shining | ref<src=tgt |
| acl:relcl | – | – | – | 0.69 | overuse of SL | ref<tgt<src |
| addit | – | – | – | 0.4 | shining | ref<src=tgt |
| adp | – | normal | – | 0.4 | useless | tgt=ref=src |
| advcl | – | – | – | 0.7 | overuse of SL | ref<tgt<src |
| advers | – | – | – | 0.39 | overuse of SL | ref<tgt<src |
| advmod | normal | – | equal | 0.53 | overuse of SL | ref<tgt<src |
| amod | – | – | – | 0.51 | russification | src<ref<tgt |
| anysome | – | – | – | 0.36 | overuse of SL | ref<tgt<src |
| appos | – | – | – | 0.39 | SL/TL-indep | tgt<ref=src |
| attrib | – | – | – | 0.51 | russification | src<ref<tgt |
| aux | – | – | – | 0.42 | overuse of SL | ref<tgt<src |
| aux:pass | – | – | equal | 0.36 | underuse of TL | tgt<ref<src |
| case | – | – | – | 0.54 | overuse of SL | ref<tgt<src |
| caus | – | – | equal | 0.36 | russification | src<ref<tgt |
| ccomp | – | – | – | 0.63 | overuse of SL | ref<tgt<src |
| cconj | – | – | – | 0.36 | adaptation | tgt=ref<src |
| compar | – | – | – | 0.38 | overuse of SL | ref<tgt<src |
| compound | – | – | equal | 0.63 | adaptation | tgt=ref<src |
| conj | – | – | – | 0.37 | SL/TL-indep | ref<tgt=src* |
| content-TTR | – | – | – | 0.36 | adaptation | src<ref=tgt |
| content-dens | – | – | equal | 0.36 | normalisation | src<tgt<ref |

| | | | | | | |
|---|---|---|---|---|---|---|
| copula | – | – | – | 0.52 | overuse of SL | ref<tgt<src |
| determ | normal | – | equal | 0.36 | overuse of SL | ref<tgt<src |
| deverbals | – | – | – | 0.45 | normalisation | src<tgt<ref |
| discourse | – | – | equal | 0.63 | useless | tgt=ref=src |
| epist | – | – | – | 0.45 | anglicisation | ref<src<tgt |
| finites | normal | – | – | 0.45 | adaptation | tgt=ref<src |
| fixed | – | – | – | 0.56 | russification | src<ref<tgt |
| flat | – | – | equal | 0.36 | underuse of TL | tgt<ref<src |
| infs | – | – | equal | 0.39 | russification | src<ref<tgt |
| interrog | – | – | – | 0.39 | shining | ref<tgt=src |
| iobj | – | – | – | 0.58 | russification | src<ref<tgt |
| mark | – | – | – | 0.74 | overuse of SL | ref<tgt<src |
| mdd | normal | – | – | 0.57 | overuse of SL | ref<tgt<src |
| mhd | – | – | equal | 0.63 | SL/TL-indep | ref=src<tgt |
| mpred | – | – | equal | 0.53 | overuse of SL | ref<tgt<src |
| mquantif | – | – | equal | 0.36 | shining | ref<src=tgt |
| neg | – | – | – | 0.41 | normalisation | src<tgt<ref |
| nmod | – | – | equal | 0.74 | adaptation | src<ref=tgt |
| nn | normal | – | equal | 0.38 | normalisation | src<tgt<ref |
| nnargs | – | – | – | 0.68 | normalisation | src<tgt<ref |
| nsubj | – | – | – | 0.82 | overuse of SL | ref<tgt<src |
| numcls | – | normal | – | 0.71 | anglicisation | ref<src<tgt |
| nummod | – | – | – | 0.46 | normalisation | src=tgt<ref |
| obj | – | – | – | 0.72 | overuse of SL | ref<tgt<src |
| obl | – | – | – | 0.64 | russification | src<ref<tgt |
| parataxis | – | – | equal | 0.64 | adaptation | src<ref=tgt |
| passives | – | – | – | 0.36 | normalisation | src<tgt<ref |
| pasttense | – | – | – | 0.51 | SL/TL-indep | tgt<ref=src |
| possp | – | – | – | 0.36 | russification | src<ref<tgt |
| ppron | – | – | – | 0.36 | overuse of SL | ref<tgt<src |
| pverbals | – | – | – | 0.36 | adaptation | src<ref=tgt |

| | | | | | | |
|---|---|---|---|---|---|---|
| sconj | – | – | – | 0.36 | russification | src<ref<tgt |
| sentlength | – | – | – | 0.61 | NA | ref<tgt |
| simple | – | – | equal | 0.67 | underuse of TL | tgt<src<ref |
| superl | – | – | equal | 0.67 | adaptation | tgt=ref<src |
| tempseq | – | – | – | 0.67 | adaptation | tgt=ref<src |
| wdlength | – | normal | – | 0.4 | NA | tgt<ref |
| xcomp | – | – | – | 0.68 | shining | ref<src=tgt |

Table F.2: Properties of UD features sorted by degree of association between values in student translations and non-translations

| feature | shapiro-tgt | shapiro-ref | bartlett | LogR-F1 | trend | means |
|---|---|---|---|---|---|---|
| acl | – | – | – | 0.56 | shining | ref<src=tgt |
| acl:relcl | – | – | – | 0.66 | overuse of SL | ref<tgt<src |
| addit | – | – | – | 0.43 | SL/TL-indep | ref=src<tgt |
| adp | – | normal | – | 0.43 | adaptation | tgt=ref<src |
| advcl | – | – | – | 0.63 | overuse of SL | ref<tgt<src |
| advers | – | – | – | 0.43 | adaptation | tgt=ref<src |
| advmod | – | – | – | 0.38 | SL/TL-indep | tgt<ref=src |
| amod | – | – | – | 0.53 | russification | src<ref<tgt |
| anysome | – | – | – | 0.43 | adaptation | tgt=ref<src |
| appos | – | – | equal | 0.37 | underuse of TL | tgt=src<ref |
| attrib | – | – | – | 0.52 | russification | src<ref<tgt |
| aux | – | – | – | 0.37 | adaptation | tgt=ref<src |
| aux:pass | – | – | – | 0.37 | adaptation | tgt=ref<src |
| case | – | – | – | 0.54 | overuse of SL | ref<tgt<src |
| caus | – | – | – | 0.37 | normalisation | src<tgt<ref |
| ccomp | – | – | – | 0.50 | overuse of SL | ref<tgt<src |
| cconj | – | – | – | 0.37 | adaptation | tgt=ref<src |
| compar | – | – | – | 0.38 | overuse of SL | ref<tgt<src |
| compound | – | – | – | 0.37 | underuse of TL | tgt<ref<src |
| conj | – | – | – | 0.46 | SL/TL-indep | ref=src<tgt |

| | | | | | | |
|---|---|---|---|---|---|---|
| content_TTR | – | – | – | 0.37 | underuse of TL | tgt<src<ref |
| content_dens | normal | – | – | 0.37 | russification | src<ref<tgt |
| copula | – | – | – | 0.51 | overuse of SL | ref<tgt<src |
| determ | – | – | – | 0.37 | overuse of SL | ref<tgt<src |
| deverbals | – | – | – | 0.45 | russification | src<ref<tgt |
| discourse | – | – | – | 0.37 | underuse of TL | tgt=src<ref |
| epist | – | – | – | 0.45 | adaptation | tgt=ref<src |
| finites | – | – | – | 0.38 | overuse of SL | ref<tgt<src |
| fixed | – | – | – | 0.58 | russification | src<ref<tgt |
| flat | – | – | – | 0.37 | underuse of TL | tgt<ref<src |
| infs | – | – | – | 0.38 | adaptation | src<ref=tgt |
| interrog | – | – | – | 0.38 | useless | tgt=ref=src |
| iobj | – | – | – | 0.37 | normalisation | src<tgt<ref |
| mark | – | – | – | 0.68 | overuse of SL | ref<tgt<src |
| mdd | – | – | – | 0.39 | overuse of SL | ref<tgt<src |
| mhd | – | – | – | 0.63 | SL/TL-indep | ref=src<tgt |
| mpred | – | – | – | 0.44 | overuse of SL | ref<tgt<src |
| mquantif | – | – | – | 0.44 | useless | tgt=ref=src |
| neg | – | – | – | 0.40 | normalisation | src<tgt<ref |
| nmod | – | – | – | 0.50 | russification | src<ref<tgt |
| nn | – | – | – | 0.37 | russification | src<ref<tgt |
| nnargs | – | – | – | 0.55 | normalisation | src<tgt<ref |
| nsubj | – | – | – | 0.71 | shining | ref<tgt=src |
| numcls | – | normal | – | 0.52 | russification | src<ref<tgt |
| nummod | – | – | – | 0.71 | adaptation | src<ref=tgt |
| obj | – | – | – | 0.61 | overuse of SL | ref<tgt<src |
| obl | – | – | – | 0.59 | russification | src<ref<tgt |
| parataxis | – | – | equal | 0.67 | normalisation | src<tgt<ref |
| passives | – | – | equal | 0.37 | normalisation | src<tgt<ref |
| pasttense | – | – | equal | 0.42 | SL/TL-indep | tgt<ref=src |
| possp | – | – | – | 0.42 | adaptation | src<ref=tgt |

| | | | | | | |
|---|---|---|---|---|---|---|
| ppron | – | – | – | 0.37 | overuse of SL | ref<tgt<src |
| pverbals | – | – | equal | 0.37 | adaptation | src<ref=tgt |
| sconj | normal | – | – | 0.37 | russification | src<ref<tgt |
| sentlength | – | – | – | 0.57 | NA | ref<tgt |
| simple | – | – | – | 0.37 | underuse of TL | tgt<ref<src |
| superl | – | – | – | 0.37 | adaptation | tgt=ref<src |
| tempseq | – | – | – | 0.37 | adaptation | tgt=ref<src |
| wdlength | normal | normal | – | 0.43 | NA | ref<tgt |
| xcomp | – | – | – | 0.52 | shining | ref<src=tgt |