Research Group in Computational Linguistics

Human Translation Quality Estimation

Empirical Translation Studies

Maria Kunilovskaya

7LN002/UM1: Corpus Linguistics with R

Wolverhampton, 4 April, 2022

Outline

- Translation Varieties and Quality
  - research basics
  - empirical study of translations
  - aspects of quality
  - approaches to annotation and learning
- Translationese and Quality Estimation
  - hand-engineered features
  - feature-learning approaches
- HowTo: A translationese study with Python
  - corpus design
  - numeric representation
  - analysis
  - explanation

Research basics

Topic

Translationese-based human translation quality estimation

Main question

Can an algorithm predict human translation quality (TQ)?
In particular:

1. How much various TQ labels and scores are aligned with various language representations?
2. Are translationese indicators useful for HTQE task?
3. How do feature-based and feature-learning approaches compare on HTQE task?

Applicability:

improving learning strategies for humans and machines

Additional curiosities about translationese

1. How much translationese is about interference?
   If deviations from the expected TL norm are established, are they:
   - SL-induced ("shining-through" effect or interference)
   - or SL/TL-independent representing a cognitive process or professional norm operating in a given register/culture/period?
2. What explains translators' choice best?
   - professionalism?
   - register/genre?
   - distance between source and target languages (ST vs TT)?
3. Which features capture (each type of) translationese best?
   - Are translationese indicators the same across genres, language pairs competence levels?

Motivation and background

Does specificity of translations correlate with perceived quality?

Translation textbooks list typical issues (for EN > RU):

1. overuse of possessive pronouns (e.g. He cut his finger);
2. lack of VSO sentences;
3. excessive analytical passives (frequency calques);
4. overuse of prepositional phrases in the absolute sentence end (e.g. I totally forgot about him);
5. abstract nouns in plural form (e.g. attitudes, struggles);
6. overuse of subordinate clauses, esp. relative (that/which);
7. a lot of *I think* and *I believe*; lack of native author's stance;
8. overuse of modal predicates; lack of parenthetical discourse markers of subjective modality;
9. overuse of connectives[21];
10. longer, wordier, more repetitive language.

Area of research: Empirical Translation Studies

'translations as texts in their own right' (early 90s)

> Empirical (Corpus-based) Translation Studies (CBTS) seeks to explain linguistic choices in translations vs. non-translations by language-pair internal or external factors.

Important factors (and translation varieties):

- translator's professionalism [25]
- register/genre [22]
- directionality and SL [12;26]
- method of translation: human vs machine

Before CBTS: targets are 'deformed' reflections of their sources

- traditional linguistic (!) TS focuses on ST/TT relations (cf. key concepts: equivalence, shifts, units, correspondences, strategies) + social/cultural impact of translations

Competence levels (professionalism) as a proxy for quality

Learner translator corpora

- output of translation education, with real-life assessment
- over 15 small-size corpus projects (1998-2021), esp. following MeLLANGE project (2005-2011)[8]
- extensive metadata, error annotation (in brat), multi-parallelism
- used for descriptive case studies

Limitations:

- no parallel error annotation for ST and TT;
- small size, heterogeneous;
- lack of consistency in real-life assessment (vs artificially controlled; experimental setups)

Russian Learner Translator Corpus: source of learner texts

- 2.3 mln wds, 4.8K texts, 26K unique source sentences
- EN>RU subcorpus: 402 sources, ≈ 8 targets for each
- 17% error-annotated (553 targets, 46 sources, 12K sentpairs)

translators

- 60% by final-year TS undergraduates (Russian L1)
- from 14 Russian universities

conditions and results

- 32% (1.5K texts) are graded
- Routine/Exam/Contest
- Class/Home

text size and genres

- RU_target size: ≈ 400 wds
- 10 genres (90% mass media)

formats and structure

- *.txt and a customised TMX
- stand-off metadata and *.ann files
- public, downloadable

Research corpus at a glance

1. subsets from Russian Learner Translator Corpus (RusLTC):

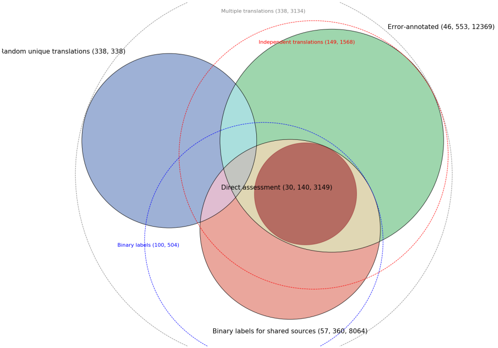


Fig.1: Number of (source, target) texts by type of annotation

2. comparable professional subcorpus: 404 sent-aligned docs (BBC Russian Service, InoSMi, RNC);
3. comparable non-translations (RNC): 8,210 > 448 docs sample

Humans vs machines: Implications for predicting quality

Translation Studies meets Computational Linguistics

How human translation (HT) differs from MT:

1. HT is essentially document-level →
   sentence-level representations less adequate
2. HT is more varied, less word-for-word →
   reference-based approaches not good
   higher granularity of quality analysis required
3. HT is expected to be 'dissemination' (publishable) quality
4. lack of reliable quality labels / available datasets →
   same as in MT: *k* 0.2-0.4 (Graham, 2015)[15]
5. no access to internal processes →
   no QuEST++ 'glassbox' features
6. HT and SOTA NMT might need to focus on different aspects of quality: fluency and accuracy respectively

What's a good translation?

How good is this translation?

Adequacy	usefulness, fitness for communicative purpose, acceptability[33;17]
Accuracy	semantic similarity: *how much of the meaning expressed in the source is also expressed in the target*
Fluency	readability, compliance to TL norms from *Flawless English* to *Incomprehensible*

Undifferentiated approach:

How much do you agree that the translation adequately expresses the meaning of the source?

Benchmarking quality by recording human judgment

(1) Real-life quality judgments:
education, certification, competitions, industrial quality control
(2) Experimental setups

Assessment purpose: quantitative or diagnostic

- summative vs formative (explanatory)
- holistic vs analytical

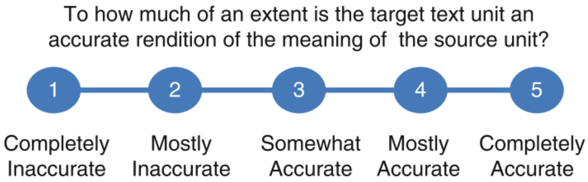
Methods:

1. direct assessment
2. (analytical) rubrics
3. errors

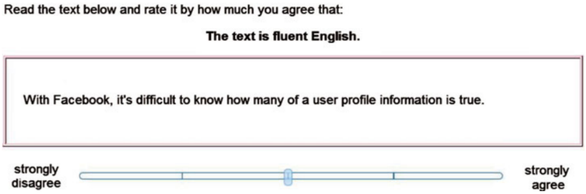
\+ in MT: post-editing time/effort (not discussed)

Granularity: document-, sentence-, word-level

Assessment method 1: Direct Assessments (DA)



from Moorkens (2018)[32]



from Graham (2015)[15]

DA: recommendations for producing MT benchmarks

from Läubli et al. (2020)[30]

- use language professionals as annotators;
- evaluate documents, not sentences; or sentences in context;
- evaluate fluency in a monolingual setup, separately;
- avoid post-editese in reference translations → use bilingual setups for accuracy;
- use original source texts (not reversed parallel corpora).

Assessment method 2: Rubrics

Diploma in Translation (DipTrans, UK certification)

1. comprehension, accuracy and register (max 50);
2. grammar, cohesion, coherence and organisation of work (max 35);
3. technical aspects: punctuation, spelling, dates, names (max 15).

BANDS: distinction, merit, pass, fail with numeric marks

American Translators Association (ATA)[46;47]

1. usefulness/transfer (max 35);
2. terminology/style (max 25);
3. idiomatic writing (max 25)
4. target mechanics (max 15)

BANDS: standard, strong, acceptable, deficient and minimal

Assessment method 3: Error annotation

e.g. harmonised DQF-MQM error taxonomy:[1]

Top-level categories (with some subcategories)

- accuracy (addition/omission, improper exact TM match, mistranslation, untranslated)
- fluency (grammar, spelling, character encoding)
- locale convention (address/currency format, shortcut key)
- style (awkward, company style, unidiomatic)
- terminology (inconsistent with termbase)
- verity (culture-specific references)

[1] https://www.qt21.eu/wp-content/uploads/2015/11/QT21-D3-1.pdf

RusLTC binary categories

'best', 'worst'

- 105 sources, inc. 57 with targets in both categories
- (but!) obtained using various scales and approaches
- random 40 triplets re-evaluated by three experts ($\alpha = 0.310$)
- after discarding 8 disputable triplets, the majority vote has the predictive accuracy of 91%, F1 = 0.912

professional varieties

(after sampling, filtering, alignment and length normalisation)

non-translations 412 docs, 32K sents, 409K words

professional 404 doc pairs, 15K sents, 320K words

students 338 doc pairs (non-multiple), 10K sents, 181K words

Continuous scores at sentence level

from error annotation

- real-life assessments 2015-2018
- based on 30 error types[24]
- IAA (for a sample) on top categories for mistakes marked in the same span $\alpha = 0.535$, 3 experts

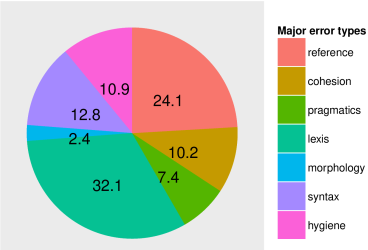
direct assessment (see blog)

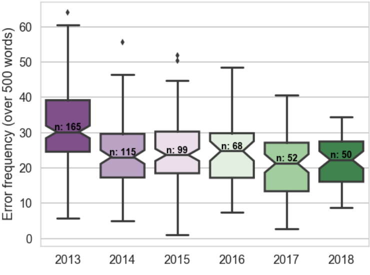
- applied rules for DA in MT[15;30] (context, slider, calibration)
- 12 final-year translation BA students
- 'adequacy' scores for 3,149 sentences (30 ST, 140 TT);
- best triple IAA:
  $\alpha = 0.303$; validity against error-based scores: r=0.257

Multiple error-based scores

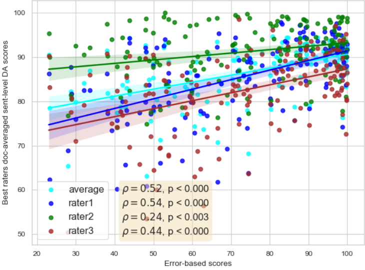
- content errors -> accuracy?
- weighted by severities (critical, major, minor)? By error type? + Kudos?
- Task: identify the most predictable score

EN > RU error type distribution





Intensity of annotation over years



Correlation: DA and error-based

Traditional approach to explaining learner language quality

Manual analysis of aligned error-tagged student translations
Top 20 most frequent triggers (real problem areas) [23]

trigger	cases
complex noun phrases	25
non-human S as agents	22
theme-rheme	21
cliché	20
nominalisations	19
terms	15
contrastive combinability	15
compression	14
complex sentences	13
SL-specific lexis	13

trigger	cases
infinitives	10
detailed descriptions	9
word order	9
polysemy/contextual	9
proper names	8
figurative speech	8
discourse markers	8
modal verbs	7
passive voice	5
plural of nouns	4

Based on 405 independent student translations to 32 English source texts, reduced to the top 5 most challenging source sentences (160 source sentences, 2K target sentences)

Quality-related NLP tasks in MT

Quality Evaluation

measure distance from a MT to another translation (aka reference), usually a human translation

Most used metrics:

- BLEU
- HTER
- ...

in HT this means punishing creativity and variety

Quality Estimation

predict quality labels without references, using

- feature-engineering: QuEst++ (Specia, 2015)[41]
- using embeddings: deepQuest[18], TransQuest[37]

Granularity:

- sentence-level
- document-level
- word-level (error detection)

Where are we?

Translation Varieties and Quality
- research basics
- empirical study of translations
- aspects of quality
- approaches to annotation and learning

Translationese and Quality Estimation
- hand-engineered features
- feature-learning approaches

HowTo: A translationese study with Python
- corpus design
- numeric representation
- analysis
- explanation

Translationese methodology

translationse: collective properties of translations that make them distinct from comparable non-translations in the TL

related tasks
- translation detection,
- SL detection,
- translation direction detection

required corpora :
- translations vs non-translation (expected TL norm)
- ideally: sent-aligned documents and register-comparable non-translations

methods
- univariate/multivariate analysis
- feature selection
- text classification
- mildly-supervised methods and exploratory clustering[12;29]

features: see below

Hypothesis of translation universals

Translation universals (originally)[2]

> *"features which typically occur in translated text rather than original utterances and which are not the result of interference from specific linguistic systems" (Baker, 1993)*[3]

- are revealed through corpus-based quantitative analysis
- describe and explain linguistic specificity of translations, *'the property of being a translation'*
- typify translations as a target language variety

Related terms:

translationese (Gellerstam, 1986)[14],
third code (Frawley, 1984)[13],
laws of translation (Toury, 1995)[44]
translational tendencies/trends, inc. interference

[1]cf. language universals such as 'Languages with dominant VSO order are always prepositional.' (Greenberg, 1963)[16]

Suggested translational tendencies (Chesterman, 2004)[9]

S-universals: properties, induced by source language (SL)

1. interference (and transfer) = 'shining through' effect[43]
   translations follow source text (ST) rather than target language (TL) patterns
   e.g. "*strange strings*"; "frequency calques": unusually low or high frequencies of TL items
2. explicitation[34]
   spelling things out rather than leave them implicit
   - more frequent use of connectives;
   - more re-phrasing, comments, elaboration in brackets;
   - ST non-finite clauses > TT finite clauses[5];
   - ST pronouns and ellipsis > TT full NPs[50]
3. levelling-out (aka Standardization/Convergence)
   translations are more homogeneous and less creative than ST
4. lengthening: *translations are longer than their sources*

Suggested translational tendencies, cont.

T-universals: properties resulting from the gravitational pull from the TL

1. simplification

   less varied vocabulary, higher readability scores, less figurative language

2. normalization

   tendency to exaggerate properties of the TL; e.g. lexical "teddy-bears"

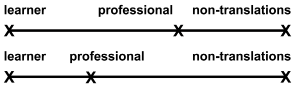
3. unique items hypothesis

   TL specific items are under-represented

NB! Matching trends and specific translationese indicators is tricky

Why use translationese for TQE

Professionals demonstrate less translationese in univariate analysis (Kunilovskaya, 2018)[27]:



The translationese-quality link is implied in:

- Scarpa (2006), Loock (2016), Sutter et al. (2017)[39;42;31;36]; for MT – Aharoni (2015), Aranberri (2020)[1;2]

Human professionalism is about fluency:

- Carl (2010)[7]: Students make more fluency errors than pro.

Fuzzy distinctions between accuracy and fluency aspects:

- Callison-Burch (2007)[6]: humans cannot differentiate aspects of quality

Translationese is a set of deviations from TL norm, i.e. disfluency

Approaches to feature engineering I

(1) Count-based features:

- frequencies of individual items/patterns (e.g. relative *that*)
- cumulative frequencies of listed items (connectives, pronouns)
- frequencies of PoS tags, syntactic dependencies (and combination)
- character[35] or word ngrams (inc. on 'mixed' representations[4])

(2) Calculated features:

- lexical variety, density, TTR
- average of senses/syllables per word
- sentence depth as parse tree depth, mean dependency distance
- ratios of N/V, 1st frequency quartile bigrams, neologisms
- Flesch Reading Ease score[38]
- LM entropy scores

Approaches to feature engineering II

(3) Embedding spaces learnt from delexicalised corpus versions:

- sequences of PoS tags, semantic tags [11;10]

Desired properties:

- well-motivated and interpretable;
- content-independent;
- reasonably frequent;
- reliably extractable;
- language-independent or shared by SL and TL

32 features from Vered Volansky (2015) [45] are used as a benchmark.
MT: QuEst++ 17/77 doc-level features for ST complexity, TT fluency and transfer adequacy (ratios of ST/TT) [Scarton].

Hand-engineered features for document-level representations

1.Structural delexicalised features from UD annotations

well-known indicators and expectations for translations:

- lexical variety, TTR (lower),
- lexical density (lower),
- overuse of discourse markers
- sentence length (higher)
- overuse of pronouns

patterns expected from English-to-Russian studies:

- mean hierarchical distance
- underuse of nsubj:pass (ex. 'resheno prodlit'), negative particles, deverbal nouns
- overuse of connectives and modal predicates

2.ngram ratios and perplexity

ratios of 1-2-3-grams from top/bottom freq quartiles
mean sentence perplexity + STD

3.collgram features

ratios of NPMI- and Tscore-score based highly and negatively collocated phrases

58 frequency features based on UD annotation, including

morphological forms

passive forms, infinitives, deverbals

morphological categories

groups of pronouns and conjunctions

UD relations

types of clauses, parataxis, auxiliary verbs

syntactic functions

copula verb, attribute, nounal subject

sentence type and structure

number of clauses, sentence length, modal predicates

graph-based features

mean dependency /hierarchical distance[19];

types of discourse markers

addit, advers, caus, tempseq, epist and but

lexical measures

lexical density and TTR based on content lemmas

Results for HTQE on translationese indicators using SVM

Translationese classification on UD-features: F1 = 0.912

Binary labels, SVM, F1-score

- best-worst: 0.635
- students-professionals: 0.733

Continuous scores, SVR, Pearson r 0.494 (document-level)

on 17 doc-level features from **QuEst++**:

F1-score over 10 folds: 0.579

Previous research in HTQE

Pearson r on continuous scores

- Yuan (2018, 2020)[47;48]
  - setting: English-to-Chinese, 458 student translations to 6 sources (sic!), 3529 sent pairs, 4 continuous scores (ATA rubrics), 360 features, e.g. freq of semantic roles tags, target source adverbial modifier log ratio + feature selection
  - best result
    - document-level (XGBoost, features): $r = 0.62\text{-}0.76$
    - sentence-level (features): $r = 0.34\text{-}0.55$
      (cf. $r = 0.31\text{-}0.41$ for CNN-BiLSTM-Att on word-vectors)
      (cf. MTQE WMT20 $r = 0.53$)
- Zhou (2019)[49]
  - setting: Japanese-to-English, unsupervised approach: correlation between ST/TT similarity/distance measures based on word vectors and overall quality graded by humans for 130 sentence pairs from camera manuals
  - result: $r = 0.53$

Attempted vector representations

Distributional models

STS for accuracy

- Embeddings capture word semantics
- Representations in two languages can be transformed into a shared semantic space (MUSE project[28])
- Cross-linguistic textual similarity (cosine) is an approximation of accuracy

perplexities for fluency

- Language models (LM) calculate the probability of a vocabulary item to be next in a sequence
- Bad (disfluent) translations have higher entropy and should result in higher LM perplexity
- Use LM perplexity from ELMo as a fluency measure

(1) Results: Binary labels, doc-level classification

best-worst (513 documents)

(human experts achieve F1 = 0.914 on a random subset)

method	F1-score
fluency (against non-translations)	
average sentence perplexities from LMs (HMM, RNN, ELMo), XGBoost	0.50-0.56
accuracy (cross-linguistic textual similarity)	
SVM, tf-idf BOW char 3-grams	0.674
SVM, concated averaged word vectors for ST&TT (from a cross-lingual model on lem-pos, no stopwords)	0.607
SVM, cosine between averaged ST&TT as a single feature	0.579
siamese BiLSTM with dot product of ST&TT sentence vectors	0.630
SVM, Quest++ 17 features	0.579

Probably not enough data for neural approaches and embeddings

(2) Results: Continuous scores from errors

converting error stats to scores is not a trivial task: 7 strategies

best option: unscaled mean for accuracy and fluency errors, regardless error types

- **doc-level (553 samples)**

	Pearson r
BiLSTM, averaged sentence vectors ELMo(lemma)	0.520

- **sent-level (12,000 samples)**

	Pearson r
BiLSTM, bag of ELMo(lemma) vectors	0.250
TransQuest (bert-base-multilingual-cased, 3 folds)	0.275

Summary and Outlook

- Results are higher for competence than for binary quality.
- The results are low, but in line with similar in the field.
- Translationese indicators are better than (averaged) word embeddings but worse than tf-idf on binary labels.
- Data refinement and better representations should yield improvement.

- wider range of indicators
- new abstract lexical features based on association measures
- SBERT
- use error annotation as input into a neural architecture

NB! Work in progress, still a lot of room for improvement

Where are we?

Translation Varieties and Quality
- research basics
- empirical study of translations
- aspects of quality
- approaches to annotation and learning

Translationese and Quality Estimation
- hand-engineered features
- feature-learning approaches

HowTo: A translationese study with Python
- corpus design
- numeric representation
- analysis
- explanation

Structured corpus

1. Re-use (and double-check) an existing resource (best for translationese: EuroParl-UdS[20])
2. Extract from XML / TMX (BNC, RNC, OPUS, RusLTC)
   - bnc_extract_media.py
   - extract-docs-from-tmx.py
   - extract-text-from-links.py and align sentences (e.g. LF Aligner)
   - googletrans_api.py
3. Build a structured corpus (use folder names as categories)
   - normalise doc size, sent length
   - check parallelism
   - annotate (lemmatise, PoS tag, parse): simple_UDparser.py

```
kateryna/corpus$ tree
├── parsed
    ├── debates
        ├── ref
            └── es
                (1031 files)
                ├── 19990721.ES.conllu
                ├── 19990722.ES.conllu
        ├── src
            └── en
                (539 files)
                ├── 19990720.EN.conllu
                ├── 19990721.EN.conllu
        └── tgt
            └── es
                (539 files)
                ├── 19990720.ES.conllu
                ├─/home/u2─ 19990721.ES.conllu

    └── fiction
        ├── ref
            └── es
                (257 files)
                ├── ref_10_chunk_11.conllu
                ├── ref_10_chunk_12.conllu
        ├── src
            └── en
                (144 files)
                ├── src_1_chunk_10.conllu
                ├── src_1_chunk_11.conllu
        └── tgt
            └── es
                (144 files)
                ├── Resulttgt_1_chunk_10.conllu
                ├── tgt_1_chunk_11.conllu

└── txt
    ├── debates/home/u2
        ├── ref
            └── es
                ├── 19990721.ES.txt
                ├── 19990722.ES.txt
        ├── src
            └── enRFE
                ├── 19990720.EN.txt
                ├── 19990721.EN.txt
        └── tgt
            └── es
                ├── 19990720.ES.txt
                ├── 19990721.ES.txt
```

Feature engineering and extraction, or vectorisation

Get a table of shape: Docs X Features

Operationalise hypotheses: put existing claims to an empirical test (Russian uses more negative sentences and more passives than English; translations have less varied vocabulary)

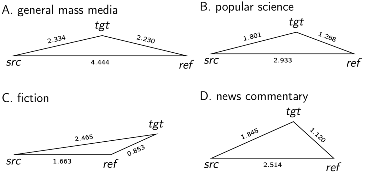
Throw a wide net: use easily extractable features and hope that you stumble upon something interesting and the results will be interpretable

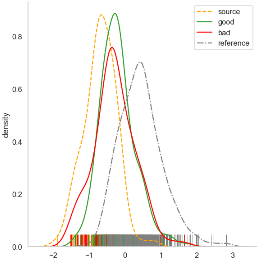
Vectorise: (as a baseline?) Beware that using surface lexical features (strings) will capture domain differences between translations and non-translation

My pipeline to extract UD-based features from *.conllu format

Univariate analysis and ML algorithms

- Do your features capture translationese? (significance tests and effect size)
- Compare frequencies in sources, targets, non-translations to establish the nature of the deviation
- Visualise differences on PCA transformed vectors
- Classify or cluster to demonstrate how good the text categories can be distinguished
- Use feature selection or internal weight analysis (ANOVA, RFE) to identify best predictors





Groups of possible explanatory factors

- contrastive studies: interference and transfer (lack of professionalism?)
- social and ethical norms (risk-minimising strategies, language prestige)
- register and professional conventions
- cognitive pressures (explicitation)

References I

[1] Aharoni, R. (2015). Automatic Detection of Machine Translated Text and Translation Quality Estimation. PhD thesis.

[2] Aranberri, N. (2020). Can translationese features help users select an MT system for post-editing? Procesamiento de Lenguaje Natural, 64(1):93–100.

[3] Baker, M. (1993). Corpus Linguistics and Translation Studies: Implications and Applications. In Text and Technology: In honour of John Sinclair, pages 232–250. J. Benjamins, Amsterdam.

[4] 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):259–274.

References II

[5] Bisiada, M. (2014). The impact of editorial guidelines on sentence splitting in german business article translations. Applied Linguistics, 37(3):354–376.

[6] Callison-Burch, C., Fordyce, C. S., Koehn, P., Monz, C., and Schroeder, J. (2007). (meta-) evaluation of machine translation. In Proceedings of the Second Workshop on Statistical Machine Translation, pages 136–158.

[7] Carl, M. and Buch-Kromann, M. (2010). Correlating translation product and translation process data of professional and student translators. 14 Annual Conference of the European Association for Machine Translation, Saint-Raphaël, France, (May).

References III

[8] 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):221–248.

[9] Chesterman, A. (2004). Hypotheses about translation universals. Claims, Changes and Challenges in Translation Studies, pages 1–14.

[10] Chowdhury, K. D., Espã na Bonet, C., van Genabith, J., and Genabith, J. (2020). Understanding Translationese in Multi-view Embedding Spaces. In Proceedings of the 28th International Conference on Computational Linguistics, pages 6056–6062, Barcelona, Spain. Online.

References IV

[11] Chowdhury, K. D., Espa˜na-Bonet, C., and van Genabith, J. (2021). Tracing Source Language Interference in Translation with Graph-Isomorphism Measures. In RANLP-2021, pages 380–390.

[12] Evert, S. and Neumann, S. (2017). The impact of translation direction on characteristics of translated texts : A multivariate analysis for English and German. In De Sutter, G., Lefer, M., and Delaere, I., editors, Empirical Translation Studies: New Methodological and Theoretical Traditions, volume 300, pages 47—-80. Walter de Gruyter GmbH & Co KG.

[13] Frawley, W. (1984). Prolegomenon to a theory of translation. Translation: Literary, Linguistic & Philosophical Perspectives, 159:175.

[14] Gellerstam, M. (1986). Translationese in Swedish novels translated from English. Translation studies in Scandinavia.

References V

[15] 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):3–30.

[16] Greenberg, J. H. (1963). Some universals of grammar with particular reference to the order of meaningful elements. Universals of language, 2:73–113.

[17] House, J. (2001). Translation Quality Assessment: Linguistic Description versus Social Evaluation. Meta: Journal des traducteurs, 46(2):243.

[18] Ive, J., Blain, F., and Specia, L. (2018). deepQuest: A Framework for Neural-based Quality Estimation. In Proceedings ofthe 27th International Conference on Computational Linguistics, pages 3146–3157.

References VI

[19] Jing, Y. and Liu, H. (2015). Mean hierarchical distance augmenting mean dependency distance. In Nivre, J. and Hajicova, E., editors, Proceedings of the Third International Conference on Dependency Linguistics (Depling 2015), pages 161–170.

[20] Karakanta, A., Vela, M., and Teich, E. (2018). Europarl-uds: Preserving and extending metadata in parliamentary debates. ParlaCLARIN: Creating and Using Parliamentary Corpora.

[21] Kunilovskaya, M. (2017). Linguistic tendencies in English to Russian translation: the case of connectives. In Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialogue 2017", volume 2, pages 221–233.

References VII

[22] Kunilovskaya, M. and Corpas Pastor, G. (2021). Translationese and register variation in English-to-Russian professional translation. In Lim, L., Li, D., and Wang, V., editors, New Perspectives on Corpus Translation Studies, New Frontiers in Translation Studies. Springer.

[23] Kunilovskaya, M., Ilyushchenya, T., Morgoun, N., and Mitkov, R. (2021a). Source language difficulties in learner translation: Evidence from an error-annotated corpus. Target.

[24] Kunilovskaya, M. and Kutuzov, A. (2015). A quantitative study of translational Russian based on a translational learner corpus. In 7th International Conference Corpus Linguistics 2015, pages 33–40. Saint-Petersburg State University.

References VIII

[25] Kunilovskaya, M. and Lapshinova-Koltunski, E. (2020). Lexicogrammatic Translationese across Two Targets and Competence Levels. In Calzolari, N., Bechet, F., Blache, P., Choukri, K., Cieri, C., Declerck, T., Goggi, S., Isahara, H., Maegaard, B., Mariani, J., and Others, A., editors, Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pages 4102–4112. The European Language Resources Association (ELRA).

[26] Kunilovskaya, M., Lapshinova-Koltunski, E., and Mitkov, R. (2021b). Translationese in Russian literary texts. In DeGaetano, S., Kazantseva, A., Reiter, N., and Szpakowicz, S., editors, Proc@bookMoorkens2018, author = Moorkens, Joss and Castilho, Sheil of the 5th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities

References IX

and Literature, pages ??–?? Association for Computational Linguistics.

[27] Kunilovskaya, M., Morgoun, N., and Pariy, A. (2018). Learner vs. professional translations into Russian: Lexical profiles. Translation and Interpreting, 10(1).

[28] Lample, G., Conneau, A., Denoyer, L., and Ranzato, M. (2017). Unsupervised machine translation using monolingual corpora only. arXiv preprint arXiv:1711.00043.

[29] Lapshinova-Koltunski, E. (2017). Exploratory analysis of dimensions influencing variation in translation. The case of text register and translation method. 300:207–234.

References X

[30] 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, 67:653–672.

[31] Loock, R. (2016). La traductologie de corpus. Presses universitaires du Septentrion, Villeneuve d'Ascq.

[32] Moorkens, J., Castilho, S., Gaspari, F., and Doherty, S. (2018). Translation Quality Assessment From Principles to Practice Machine Tr Springer International Publishing AG.

[33] Nord, C. (2006). Translating as a purposeful activity: a prospective approach. TEFLIN Journal: A publication on the teaching and . . . , 17(2):131–143.

[34] Olohan, M. (2001). Spelling out the optionals in translation : a corpus study. UCREL Technical Papers, (13):423–432.

References XI

[35] Popescu, M. (2011). Studying Translationese at the Character Level. Proceedings of the International Conference Recent Advances in Natural Language Processing 2011, (September):634–639.

[36] Rabadán, R., Labrador, B., and Ramón, N. (2009). Corpus-based contrastive analysis and translation universals A tool for translation quality assessment. Babel, 55(4):303–328.

[37] Ranasinghe, T., Orasan, C., and Mitkov, R. (2020). TransQuest at WMT2020: Sentence-Level Direct Assessment. In Proceedings ofthe 5th Conference on Machine Translation (WMT), pages 1049–1055.

References XII

[38] 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):293–325.

[39] 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., editors, Insights into specialized translation, pages 155–172. Peter Lang, Bern.

[Scarton] Scarton, C. Document-Level Machine Translation Quality Estimation. (September).

References XIII

[41] Specia, L., Paetzold, G. H., and Scarton, C. (2015). Multi-level translation quality prediction with QUEST++. In Chen, H.-H. and Markert, K., editors, ACL-IJCNLP 2015 - 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing, Proceedings of System Demonstrations, pages 115–120. Association for Computational Linguistics.

[42] Sutter, G. D., Cappelle, B., and Loock, R. (2017). Towards a corpus-based, statistical approach to translation quality: Measuring and visualizing linguistic deviance in student translation. Linguistica Antverpiensia, New Series, pages 25–39.

[43] Teich, E. (2003). Cross-linguistic variation in system and text. A methodology for the investigation of translations and comparable texts. Walter de Gruyter, Berlin/Boston.

References XIV

[44] Toury, G. (1995). Descriptive Translation Studies - and Beyond. John Benjamins Publishing Company, benjamins translation library edition.

[45] Volansky, V., Ordan, N., and Wintner, S. (2015). On the features of translationese. Digital Scholarship in the Humanities, 30(1):98–118.

[46] Williams, M. (2013). A holistic-componential model for assessing translation student performance and competency. Mutatis Mutandis, 6(2):419–443.

[47] Yuan, Y. (2018). Human Translation Quality Estimation: Feature-based and Deep Learni
PhD thesis, University of Leeds.

References XV

[48] Yuan, Y. and Sharoff, S. (2020). Sentence Level Human Translation Quality Estimation with Attention-based Neural Networks. In Proceedings of the 12th Language Resources and Evaluation Conference, pages 1858–1865.

[49] Zhou, Y. and Bollegala, D. (2019). Unsupervised Evaluation of Human Translation Quality. In Fred, Ana and Filipe, J., editor, Proceedings of the 11th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management, Vienna. SCITEPRESS-Science and Technology Publications.

[50] Zinsmeister, H., Dipper, S., and Seiss, M. (2012). Abstract pronominal anaphors and label nouns in german and english: Selected case studies and quantitative investigations. Translation: Computation, Corpora, Cognition, 2(1).

Example: deverbals and relative clauses

When we **assess** how a **changing** planet could **affect** us, let's take a lesson from the Egyptians.

1. И когда мы оцениваем то, как меняющаяся планета могла бы повлиять на нас, давайте брать урок у египтян. [*And when we **assess** (that), how the **changing** planet could **influence** us, let's take a lesson from the Egyptians.*]
2. Когда мы поймём, какое влияние оказывают на нас происходящие на Земле изменения, следует вспомнить уроки, которые преподала жизнь египтянам. [*And when we **understand** what **influence is exerted** upon us by the **changes (happening)** on the Earth, we should remember the lessons, which life taught to the Egyptians*]
3. Не стоит забывать о судьбе древних египтян при оценке возможных последствий любых изменений климата на нашей планете! [*It's worth not to forget about the destiny of the ancient Egyptians at the **evaluation** of the possible **consequences** of any **changes** of climate on our planet.*]