Vilém Zouhar

vilem.zouhar@gmail.com · vilda.net

Education

ETH Zürich 2022-present
PhD in Computer Science in **LRE lab**
Advisors: Mrinmaya Sachan, Menna El-Assady
Thesis: Prudent NLG Evaluation

Saarland University + Groningen University 2020 - 2022
MSc. in Language Science and Technology (double)
Advisors: Dietrich Klakow, Gosse Bouma
Thesis **Shrinking Knowledge Base Size: Dimension Reduction & Filtering**

Charles University in Prague 2017 - 2020
BSc. in Computer Science (computational linguistics minor)
Advisor: Ondřej Bojar
Thesis **Enabling Outbound Machine Translation**

Research Interests (inter alia)

- Efficient NLP evaluation — *How to compare models robustly yet economically?*
- Non-mainstream MT and other NLG — *How to reliably assess the multilingual model quality?*
- NLP-oriented human-computer interaction — *How to convey model's confidence & quality to users?*

Work Experience

- **Google Translate**, Mountain View (research intern, evaluation) 2025
- Microsoft Translate, remote (applied scientist intern, model evaluation) 2024
- **Amazon Machine Translation**, New York (applied scientist intern, quality estimation) 2023
- **Spoken Language Systems**, Saarland University (student research assistant) 2021
- **Institute of Formal and Applied Linguistics**, Charles University (student research assist.) 2019-2022
- Previo (software dev intern) 2018
- BIM Project (software dev intern) 2017

Awards and Scholarships

- Cohere Labs Catalyst grant for efficient evaluation 2026
- Google PhD Fellowship in Natural Language Processing 2025
- Two EAMT grants for research on compute-efficient evaluation, and LLM brittleness 2025, 2025
- Language & Communication Technologies full MSc Erasmus Mundus **scholarship** 2020-2022
- Student faculty grant for **learning materials** for Machine Learning Excercises in R 2018

Teaching

- Student projects supervision 2023-present
- Deep Learning TA **2025**
- Interactive Machine Learning Visualization and Explainability TA **2025**
- Large Language Models TA **2023**, **2024**, **2025**
- Neural Networks Implementation and Applications head TA **2021**
- Statistical Natural Language Processing head TA **2021**

Selected Publications

Google Scholar Semantic Scholar

Pearmut: Human Evaluation of Translation Made Trivial (in review 2026)
Vilém Zouhar, Tom Kocmi

Estimating Machine Translation Difficulty (EMNLP 2025)
Lorenzo Proietti,[=] Stefano Perrella,[=] **Vilém Zouhar**,[=] Roberto Navigli, Tom Kocmi

Findings of the WMT25 general machine translation shared task: Time to stop evaluating on easy test sets (WMT 2025)
Tom Kocmi, **V. Zouhar** and others

Searching for Difficult-to-Translate Test Examples at Scale (In review 2026)
Wenda Xu,[=] **Vilém Zouhar**,[=] Parker Riley, Mara Finkelstein, Markus Freitag, Daniel Deutsch

Pitfalls and Outlooks in Using COMET (WMT 2024)
Vilém Zouhar,[=] Pinzhen Chen,[=] Tsz Kin Lam, Nikita Moghe, Barry Haddow

Fine-Tuned Machine Translation Metrics Struggle in Unseen Domains (ACL 2024)
Vilém Zouhar, Shuoyang Ding, Anna Currey, Tatyana Badeka, Jenyuan Wang, Brian Thompson

WMT24 General Machine Translation Shared Task: The LLM Era is Here but MT is Not Solved Yet (WMT 2024)
Tom Kocmi, **V. Zouhar** and others

RELIC: Investigating Large Language Model Responses using Self-Consistency (CHI 2024)
Furui Cheng, **Vilém Zouhar**, Simran Arora, Mrinmaya Sachan, Hendrik Strobelt, Mennatallah El-Assady

Evaluating Optimal Reference Translations (JNLE 2024)
Vilém Zouhar, Věra Kloudová, Martin Popel, Ondřej Bojar

WMT 2023 Shared Task on Machine Translation with Terminologies (EMNLP 2023)
Kirill Semenov, **Vilém Zouhar**, Tom Kocmi, Dongdong Zhang, Wangchunshu Zhou, Yuchen Eleanor Jiang

A Formal Perspective on Byte-Pair Encoding (ACL 2023)
Vilém Zouhar, Clara Meister, Juan Luis Gastaldi, Li Du, Tim Vieira, Mrinmaya Sachan, Ryan Cotterell

Poor Man's Quality Estimation: Predicting Ref.-Based MT Metrics Without Reference (EACL 2023)
Vilém Zouhar, Shehzaad Dhuliawala, Wangchunshu Zhou, Nico Daheim, Tom Kocmi, Yuchen Eleanor Jiang, Mrinmaya Sachan

How to Select Datapoints for Efficient Human Evaluation of NLG Models? (TACL 2025)
Vilém Zouhar, Peng Cui, Mrinmaya Sachan

Early-Exit and Instant Confidence Translation Quality Estimation (EACL 2026)
Vilém Zouhar, Maike Züfle, Beni Egressy, Julius Cheng, Jan Niehues

Generating Difficult-to-Translate Texts (In review 2026)
Vilém Zouhar, Wenda Xu, Parker Riley, Juraj Juraska, Mara Finkelstein, Markus Freitag, Dan Deutsch

AI-Assisted Human Evaluation of Machine Translation (NAACL 2025)
Vilém Zouhar,[=] Tom Kocmi,[=] Mrinmaya Sachan

Error Span Annotation: A Balanced Approach for Human Evaluation of Machine Translation (WMT 2024)
Tom Kocmi,[=] **Vilém Zouhar**,[=] Eleftherios Avramidis, Roman Grundkiewicz, Marzena Karpinska, Maja Popović, Mrinmaya Sachan, Mariya Shmatova

Quality and Quantity of Machine Translation References for Automated Metrics (HumEval 2024)
Vilém Zouhar, Ondřej Bojar

Navigating the Metrics Maze: Reconciling Score Magnitudes and Accuracies (ACL 2024)
Tom Kocmi, **Vilém Zouhar**, Christian Federmann, Matt Post

Distributional Properties of Subword Regularization (EMNLP 2024)
Marco Cognetta, **Vilém Zouhar**, Naoaki Okazaki

A Diachronic Perspective on User Trust in AI under Uncertainty (EMNLP 2023)
Shehzaad Dhuliawala,[=] **Vilém Zouhar**,[=] Mennatallah El-Assady, Mrinmaya Sachan

Tokenization and the Noiseless Channel (ACL 2023)
Vilém Zouhar, Clara Meister, Juan Luis Gastaldi, Li Du, Mrinmaya Sachan, Ryan Cotterell

Re-visiting Automated Topic Model Evaluation with Large Language Models (EMNLP 2023)
Dominik Stammbach, **Vilém Zouhar**, Alexander Hoyle, Mrinmaya Sachan, Elliott Ash

Neural Machine Translation Quality and Post-Editing Performance (EMNLP 2021)
Vilém Zouhar, Ondřej Bojar, Martin Popel, Aleš Tamchyna

Providing Backtranslation Improves Users Confidence in MT, Not Quality (NAACL 2021)
V. Zouhar, M. Novák, M. Žilinec, O. Bojar, M. Obregón, R. L. Hill, F. Blain, M. Fomicheva, L. Specia, L. Yankovskaya

WMT20 Document-Level Markable Error Exploration (WMT 2020)
Vilém Zouhar, Tereza Vojtěchová, Ondřej Bojar

Other Projects

GitHub

How Important is `Perfect' English for Machine Translation Prompts? (EACL 2026)

Hearing to Translate: The Effectiveness of Speech Modality Integration into LLMs (In review 2026)

Findings of the WMT25 shared task on automated translation evaluation systems: Linguistic diversity is challenging and references still help (WMT 2025)

COMET-poly: Machine Translation Metric Grounded in Other Candidates (WMT 2025)

Co-DETECT: Collaborative Discovery of Edge Cases in Text Classification (EMNLP 2025 Demo)

Biased Tales: Cultural and Topic Bias in Generating Children's Stories (EMNLP 2025)

QE4PE: Word-level Quality Estimation for Human Post-Editing (TACL 2025)

Findings of the IWSLT 2025 Evaluation Campaign (IWSLT 2025)

Interactive Analysis of LLMs using Meaningful Counterfactuals (IEEEvis 2025)

Two Counterexamples to Tokenization and the Noiseless Channel (LREC-COLING 2024)

Sampling and Filtering of Neural Machine Translation Distillation Data (NAACL SRW 2021)

How to Engage Your Readers? Generating Guiding Questions to Promote Active Reading (ACL 2024)

Enhancing Textbooks with Visuals from the Web for Improved Learning (EMNLP 2023)

Leveraging Neural Machine Translation for Word Alignment (PBML 116)

Can Large Language Models Capture Human Annotator Disagreements? (EACL 2026)

Findings of the WMT25 Multilingual Instruction Shared Task: Persistent Hurdles in Reasoning, Generation, and Evaluation (WMT 2025)

Findings of the WMT25 Terminology Translation Task: Terminology is Useful Especially for Good MTs (WMT 2025)

Deconstructing Self-Bias in LLM-generated Translation Benchmarks (In review 2026)

Unsupervised Word-level Quality Estimation for Machine Translation Through the Lens of Annotators (Dis)agreement (EMNLP 2025)

Large Language Models as Span Annotators (In review 2026)

A Bayesian Optimization Approach to Machine Translation Reranking (NAACL 2025)

Are Large Language Models for Education Reliable for All Languages? (BEA 2025)

PWESuite: Phonetic Word Embeddings and Tasks They Facilitate (LREC-COLING 2024)

Knowledge Base Index Compression via Dimensionality and Precision Reduction (SpaNLP 2022)

Harmonizing Assistance: Moderating Visual andTextual Aids in AI-Enhanced Textbook Readingwith IRead (IJAIED 2025)

AutoTutor meets Large Language Models: A Language Model Tutor with Rich Pedagogy and Guardrails (Learning@Scale 2024)

Artefact Retrieval: Overview of NLP Models with Knowledge Base Access (AKBC CSKB 2021)

Sentence Ambiguity, Grammaticality and Complexity Probes (BlackboxNLP 2022)

Invited talks

- Selecting Examples to Human-Evaluate NLG Models at Charles University (2025), TU Darmstadt (2025), UMD Multilingual Seminar (2026), SICSA (2026)
- Prudent NLG Evaluation at KIT (2025), Cardiff NLP Seminars (**2025**), Google Translate (2024)
- Token(s) of Appreciation for BPE at MT Marathon (**2024**)
- ESA and ESAAI at Microsoft Translate (2024)
- How we solved tokenization but got it wrong at ZurichNLP Meetup #9 (**2024**)
- Quality and Quantity of Machine Translation References for Automated Metrics at IST/Unbabel seminar (**2024**)
- Poor Man's Quality Estimation at IST/Unbabel seminar (**2023**)

Service

- **Multilingual Multicultural Evaluation Workshop** at EACL (organizer) 2025
- **Multilingual Instruction** Shared Task at WMT (co-organizer) 2025
- **Evaluation Metrics** Shared Task at WMT (co-organizer) 2025
- **General Machine Translation** Shared Task at WMT (co-organizer) 2024, 2025
- **WMT Terminology** Shared Task (organizer) 2023, 2025
- Evaluation committee member for granting university accreditations in Czechia 2020-2025
- **CSRR**: Workshop on Commonsense Representation and Reasoning (organizer) ACL 2022
- Reviewer: {EACL, MME, ACL, SwissText}2026
  {NAACL★, ACL, EMNLP, WMT, ARR}2025
  {LREC-COLING, EACL★, ARR, Repl4NLP, EMNLP, WMT}2024
  {ARR, WMT, EMNLP, AACL-IJCNLP, ACL}2023
  {SVRHM, CoNLL, AACL-IJCNLP}2022

Technical Knowledge

GitHub

- **Programming:** Python, JS/TS, Rust
- **Toolkits:** PyTorch, Scikit, Numpy, HF, Fairseq, Marian NMT
- **Misc:** Linux (long-term user, cluster), passionate about visualization, writing and typesetting (Typst, LaTeX)

Languages

- Czech: C2
- English: C2
- German: B2

+ linguistic curiosity

Students

Pleasure to (co-)supervise student projects/theses

- Rosamund Ang: Sample-Efficient Aligning of Automatic Metrics ETH 2026
- Šimon Sukup, William Kalikman, Michal Tešnar: Augmenting Texts to Increase Translation Difficulty ETH 2026
- Victor Zarzu: Speech Translation Quality Estimation ETH 2026
- Chenfei Xiong: Collaborative Discovery of Edge Cases in Text Classification ETH 2025
- Oymak Gül, Hatipoglu Öykü, Garries Urbina Oscar, Victor Zarzu Interactive Visualization of Chain-of-Thought Reasoning in Large Language Models ETH 2025
- Yijie Tong, Haokun He (Self Quality Estimation for Machine Translation and NLP) UZH 2024
- Abhinav Kumar (Data Augmentation for Text Complexity Prediction) UZH 2024
- Vincent Dörig (IRead: AI-Enhanced Textbook Reading) ETH 2024
- David Gu (Manifestations of Image Complexity) ETH 2023

Extra-Curricular

- Academic senate at Charles University Faculty of Mathematics and Physics 2018-2020
- Several **games** programmed and presented in limited time, mostly Ludum Dare 2015-present
- Organization of **Kasiopea**, an annual coding competition for talented high school students 2017-2019
- Amateur electric guitar  2021-present