

# 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: Quality- and Complexity-Aware NLP Evaluation



UNIVERSITÄT  
DES  
SAARLANDES



university of  
 groningen

### Saarland University + Groningen University

(2 years) 2020 - 2022

MSc. in Language Science and Technology (double)

Advisors: Dietrich Klakow, Gosse Bouma

Thesis **Shrinking Knowledge Base Size: Dimension Reduction, Splitting & Filtering**



### Charles University in Prague

(3 years) 2017 - 2020

BSc. in Computer Science (computational linguistics minor)

Advisor: Ondřej Bojar

Thesis **Enabling Outbound Machine Translation**



## RESEARCH INTERESTS (inter alia)

- ▶ Non-mainstream MT and other NLP *How to reliably predict the quality of MT (and other NLP) output?*
- ▶ Efficient NLP evaluation *Which examples to choose to efficiently compare a set of models?*
- ▶ NLP-oriented human-computer interaction *How to convey model's confidence to the users to increase trust?*



## TEACHING

- ▶ Student projects supervision 2023-present
- ▶ Large Language Models tutor 2023, 2024
- ▶ Neural Networks Implementation and Applications **tutoring & materials** winter semester 2021
- ▶ Statistical Natural Language Processing **tutoring & materials** summer semester 2021



## WORK EXPERIENCE

- ▶ **Amazon Machine Translation**, New York (applied scientist intern, quality estimation) (4 months) 2023
- ▶ **Spoken Language Systems** group (student research assistant) (1 year) 2021-2022
  - ▶ Saarland University, Germany
  - ▶ Information retrieval and language modelling efficiency
- ▶ **Institute of Formal and Applied Linguistics** (student research assistant) (3 years) 2019-2022
  - ▶ Charles University, Czech Republic
  - ▶ MT- and psycholinguistics-related projects (**Bergamot**/in-browser MT, **ELITR**)
- ▶ Previo (software dev intern) (3 months) 2018
- ▶ BIM Project (software dev intern) (3 months) 2017



## AWARDS AND SCHOLARSHIPS

- ▶ Language & Communication Technologies full MSc Erasmus Mundus **scholarship** 2020-2022 (2 years)
- ▶ Student faculty grant for **learning materials** for Machine Learning Exercises in R 2018

### Fine-Tuned Machine Translation Metrics Struggle in Unseen Domains (ACL 2024)

**Vilém Zouhar**, Shuoyang Ding, Anna Currey, Tatyana Badeka, Jenyuan Wang, Brian Thompson

### RELIC: Investigating Large Language Model Responses using Self-Consistency (CHI 2024)

Furui Cheng, **Vilém Zouhar**, Simran Arora, Mrinmaya Sachan, Hendrik Strobelt, Mennatallah El-Assady

### WMT 2023 Shared Task on Machine Translation with Terminologies (EMNLP 2023)

Kirill Semenov, **Vilém Zouhar**, Tom Kocmi, Dongdong Zhang, Wangchunshu Zhou, Yuchen Eleanor Jiang

### Tokenization and the Noiseless Channel (ACL 2023)

**Vilém Zouhar**, Clara Meister, Juan Luis Gastaldi, Li Du, Mrinmaya Sachan, Ryan Cotterell

### Evaluating Optimal Reference Translations (JNLE 2024)

**Vilém Zouhar**, Věra Kloudová, Martin Popel, Ondřej Bojar

### Interactive Analysis of LLMs using Meaningful Counterfactuals (In review 2024)

Furui Cheng, **Vilém Zouhar**, Robin Shing Moon Chan, Daniel Fürst, Hendrik Strobelt, Mennatallah El-Assady

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

**Vilém Zouhar**, Calvin Chang, Chenxuan Cui, Nathaniel Carlson, Nathaniel Robinson, Mrinmaya Sachan, David Mortensen

### Scaling the Authoring of AutoTutors with Large Language Models (Learning@Scale 2024)

Sankalan Pal Chowdhury, **Vilém Zouhar**, Mrinmaya Sachan

### Sentence Ambiguity, Grammaticality and Complexity Probes (BlackboxNLP 2022)

Sunit Bhattacharya, **Vilém Zouhar**, Ondřej Bojar

### Neural Machine Translation Quality and Post-Editing Performance (EMNLP 2021)

**Vilém Zouhar**, Ondřej Bojar, Martin Popel, Aleš Tamchyna

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

**Vilém Zouhar**, Marius Mosbach, Debanjali Biswas, Dietrich Klakow

### Quality and Quantity of Machine Translation References for Automated Metrics (HumEval 2024)

**Vilém Zouhar**, Ondřej Bojar

### A Diachronic Perspective on User Trust in AI under Uncertainty (EMNLP 2023)

Shehzaad Dhuliawala, **Vilém Zouhar**, Mennatallah El-Assady, Mrinmaya Sachan

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

Janvijay Singh, **Vilém Zouhar**, Mrinmaya Sachan

### A Formal Perspective on Byte-Pair Encoding (ACL 2023)

**Vilém Zouhar**, Clara Meister, Juan Luis Gastaldi, Li Du, Tim Vieira, 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

### Navigating the Metrics Maze: Reconciling Score Magnitudes and Accuracies (ACL 2024)

Tom Kocmi, **Vilém Zouhar**, Christian Federmann, Matt Post

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

Marco Cignetta, **Vilém Zouhar**, Sangwhan Moon, Naoaki Okazaki

### 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

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

**Vilém Zouhar**, Marius Mosbach, Miaoran Zhang, Dietrich Klakow

### Providing Backtranslation Improves Users Confidence in MT, Not Quality (NAACL 2021)

V. Zouhar, M. Novák, M. Žilinc, O. Bojar, M. Obregón, R. L. Hill, F. Blain, M. Fomicheva, L. Specia, L. Yankovskaya

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

**Vilém Zouhar**

## Leveraging Neural Machine Translation for Word Alignment (PBML 116)

Vilém Zouhar, Daria Pylypenko

## Extending Ptakopět for MT User Interaction Experiments (PBML 115)

Vilém Zouhar, Michal Novák

## WMT20 Document-Level Markable Error Exploration (WMT 2020)

Vilém Zouhar, Tereza Vojtěchová, Ondřej Bojar

## Outbound Translation User Interface Ptakopět: A Pilot Study (LREC 2020)

Vilém Zouhar, Ondřej Bojar

## MISC. PROJECTS

GitHub

Machine Translation that Peeks at the Reference (2023)

Ryanize bib (2023)

Multimodal Shannon Game with Images (2022)

Machine Translate (2022)

Poetry, Songs, Literature, Legalese and Translationese (2023)

SlowAlign (2020)

Metaphor Preservation in Machine Translation and Paraphrasing (2023)

Stolen Subwords (2023)

ÚFAL Bilingual scientific abstracts corpus (2022)

Random Strum Pattern Generator (2022)

EMMT: An eye-tracking, EEG and audio corpus for multi-modal reading and translation (2021)

Slow Align Displayer (2020)

## SERVICE

- ▶ Shared Task on Subword Tokenization (co-organizer) **SIGMORPHON 2024**
- ▶ Conference on Machine Translation WMT (co-organizer) 2024
- ▶ **WMT Terminology** Shared Task (organizer) 2023
- ▶ Evaluation committee member for granting university accreditations in Czechia (5 times) 2020-present
- ▶ **CSRR**: Workshop on Commonsense Representation and Reasoning (organizer) ACL 2022
- ▶ Reviewer: {LREC-COLING, EACL★, ARR, Repl4NLP} 2024  
{ARR, WMT, EMNLP, AACL-IJCNLP, ACL} 2023  
{SVRHM, CoNLL, AACL-IJCNLP} 2022

## TECHNICAL KNOWLEDGE

GitHub

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

## LANGUAGES


- ▶ Czech: C2
- ▶ English: C2
- ▶ German: B2  
+ linguistic curiosity

## STUDENTS

Pleasure to (co-)supervise student projects/theses

- ▶ 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