

VILÉM ZOUHAR

vilem.zouhar@gmail.com · vilda.net

EDUCATION



ETH Zürich

2022-present

PhD in Computer Science in **NLPED lab**
Advisors: Mrinmaya Sachan, Menna El-Assady
Thesis: **Quality- and Complexity-Aware NLP for Humans**



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 (comp. linguistics minor)
Advisor: Ondřej Bojar
Thesis **Enabling Outbound Machine Translation**

RESEARCH INTERESTS (inter alia)

- ▶ NLP-oriented human-computer interaction *How to convey model's confidence to the users to increase trust?*
- ▶ Non-mainstream MT and other NLP *How to reliably predict the quality of MT (and other NLP) output?*
- ▶ Text complexity and simplification *How to estimate the perceived quality and complexity of text by user?*

TEACHING

- ▶ Student projects supervision 2023-present
- ▶ Large Language Models **materials contribution** summer semester 2023
- ▶ 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

STUDENTS

Pleasure to supervise student projects/theses

- ▶ Yijie Tong, Haokun He (Self Quality Estimation for Machine Translation and NLP) UZH 2023
- ▶ Abhinav Kumar (Data Augmentation for Text Complexity Prediction) UZH 2023
- ▶ David Gu (Manifestations of Image Complexity) ETH 2023

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

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

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

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

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

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

PWESuite: Phonetic Word Embeddings and Tasks They Facilitate (In review 2024)
Vilém Zouhar, Calvin Chang, Chenxuan Cui, Nathaniel Carlson, Nathaniel Robinson, Mrinmaya Sachan, David Mortensen

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

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

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

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

Re-visiting Automated Topic Model Evaluation with Large Language Models (EMNLP 2023)

 Dominik Stambach, **Vilém Zouhar**, Alexander Hoyle, Mrinmaya Sachan, Elliott Ash

Evaluating Optimal Reference Translations

(JNLE, to appear 2024)

Vilém Zouhar, Věra Kloudová, Martin Popel, Ondřej Bojar

RELIC: Investigating Large Language Model Responses using Self-Consistency (In review 2024)

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

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. Žilínek, 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
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

Quality and Quantity of Machine Translation References for Automated Metrics (In review 2024)
Ryanize bib (2023)

Tool to check for common BibTeX best practice violations

Metaphor Preservation in Machine Translation and Paraphrasing (2023)
Poetry, Songs, Literature, Legalese and Translationese (2023)

Essay on automated sentence complexity perspective

Stolen Subwords (2023)

Report on importance of vocabularies for machine translation model stealing

Bilingual scientific abstracts corpus (2022)

Bilingual Czech and English abstracts of ÚFAL papers. With Rudolf Rosa.

Machine Translate (2022)

Open resources and community for machine translation (contributor)

Dorfromantik solver (2023)

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

With Sunit Bhattacharya, Věra Kloudová, Ondřej Bojar

Fact Learning with Adaptive Color Palette: Effect of Stimuli-Independent Hints (2021)

With Leander van Boven, Tianyi Li and Anjali Nair.

MosQEto (2019)

Quality estimation data augmentation. With Ondřej Měkota.

Slow Align Displayer (2020)

Creates quick graphs given word alignment.

ZimaDB (2018)

SQLite-like database implementation. With Petr Chmel.

SMAKE (2019)

Simple Markable And Keyword Extraction. With Petr Houška.

Prolog KNN (2017)

Implementation of kNN in Prolog, a language not suited for this.



SERVICE

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



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



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

Multimodal Shannon Game with Images

(Preprint 2022)

With Sunit Bhattacharya, Ondřej Bojar

Stroop Effect in Multi-Modal Sight Translation

(Preprint 2022)

With Sunit Bhattacharya, Věra Kloudová, Ondřej Bojar

Random Strum Pattern Generator (2022)

Fusing Sentence Embeddings Into LSTM-based Autoregressive Language Models (Preprint 2021)

With Marius Mosbach, Dietrich Klakow

Deep Molecule QSPR (2021)

Predicting key temperature points (e.g. boiling point) of molecules.

With Nikola Kalábová.

Hyperparameters of RNN for POS Tagging using Surface-Level BERT Embeddings (2020)

SlowAlign (2020)

IBM model-based word aligned with extra features and heuristics.

Call for Menza (2019)

Daily menus around Charles University. Unmaintained.

TNTranslator (2019)

Translation inspector for n-best list navigator.

ASM Hell (2017)

Learn basic assembly through a game (LD41 submission).