

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

SERVICE

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



- **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
- **Evaluating Optimal Reference Translations** JNLE, to appear 2023
 Vilém Zouhar, Věra Kloudová, Martin Popel, Ondřej Bojar
- **Re-visiting Automated Topic Model Evaluation with Large Language Models** In review 2023
 Dominik Stambach, Vilém Zouhar, Alexander Hoyle, Mrinmaya Sachan, Elliott Ash
- **Enhancing Textbooks with Visuals from the Web for Improved Learning** In review 2023
 Janvijay Singh, Vilém Zouhar, Mrinmaya Sachan
- **PWESuite: Phonetic Word Embeddings and Tasks They Facilitate** In review 2023
 Vilém Zouhar, Kalvin Chang, Chenxuan Cui, Nathaniel Carlson, Nathaniel Robinson, Mrinmaya Sachan, David Mortensen
- **A Diachronic Perspective on User Trust in AI under Uncertainty** In review 2023
 Shehzaad Zuzar Dhuliawala, Vilém Zouhar, Mennatallah El-Assady, Mrinmaya Sachan
- **Poor Man's Quality Estimation: Predicting Ref,-Based MT Metrics Without the Reference** EACL 2023
 Vilém Zouhar, Shehzaad Dhuliawala, Wangchunshu Zhou, Nico Daheim, Tom Kocmi, Yuchen Eleanor Jiang, Mrinmaya Sachan
- **Sentence Ambiguity, Grammaticality and Complexity Probes** BlackboxNLP 2022
 Sunit Bhattacharya, Vilém Zouhar, Ondřej Bojar
- **Knowledge Base Index Compression via Dimensionality and Precision Reduction** ACL Spa-NLP 2022
 Vilém Zouhar, Marius Mosbach, Miaoran Zhang, Dietrich Klakow
- **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. Žilínek, O. Bojar, M. Obregón, R. L. Hill, F. Blain, M. Fomicheva, L. Specia, L. Yankovskaya
- **Artefact Retrieval: Overview of NLP Models with Knowledge Base Access** AKBC CSKB 2021
 Vilém Zouhar, Marius Mosbach, Debanjali Biswas, Dietrich Klakow
- **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
- **WMT20 Document-Level Markable Error Exploration** WMT 2020
 Vilém Zouhar, Tereza Vojtěchová, Ondřej Bojar
- **Extending Ptakopět for MT User Interaction Experiments** PBML 115
 Vilém Zouhar, Michal Novák
- **Outbound Translation User Interface Ptakopět: A Pilot Study** LREC 2020
 Vilém Zouhar, Ondřej Bojar

► Metaphor Preservation in Machine Translation and Paraphrasing	2023
► Ryanize bib Tool to check for common BibTeX best practice violations	2023
► Poetry, Songs, Literature, Legalese and Translationese Essay on automated sentence complexity perspective	2023
► Stolen Subwords Report on importance of vocabularies for machine translation model stealing	2023
► Multimodal Shannon Game with Images With Sunit Bhattacharya, Ondřej Bojar	Preprint 2022
► Bilingual scientific abstracts corpus Bilingual Czech and English abstracts of ÚFAL papers. With Rudolf Rosa.	2022
► Stroop Effect in Multi-Modal Sight Translation With Sunit Bhattacharya, Věra Kloudová, Ondřej Bojar	Preprint 2022
► Machine Translate Open resources and community for machine translation (contributor)	2022
► Random Strum Pattern Generator	2022
► Fusing Sentence Embeddings Into LSTM-based Autoregressive Language Models With Marius Mosbach, Dietrich Klakow	Preprint 2021
► EMMT: An eye-tracking, EEG and audio corpus for multi-modal reading and translation With Sunit Bhattacharya, Věra Kloudová, Ondřej Bojar	Preprint 2021
► Deep Molecule QSPR Predicting key temperature points (e.g. boiling point) of molecules. With Nikola Kalábová.	2021
► Fact Learning with Adaptive Color Palette: Effect of Stimuli-Independent Hints Enhancing fact learning with colors. With Leander van Boven, Tianyi Li and Anjali Nair.	2021
► Hyperparameters of RNN Architectures for POS Tagging using Surface-Level BERT Embeddings	2020
► SlowAlign IBM model-based word aligned with extra features and heuristics.	2020
► Slow Align Displayer Creates quick graphs given word alignment.	2020
► MosQEto Synthesising machine translation quality estimation data. With Ondřej Měkota.	2019
► Dorfromantik solver	2023
► Call for Menza Aggregator of daily menus around Charles University. Unmaintained since I moved out of Prague.	2019
► SMAKE Simple Markable And Keyword Extraction. With Petr Houška.	2019
► TNTranslator Translation inspector for n-best list navigator.	2019
► ZimaDB SQLite-like database implementation from scratch. With Petr Chmel.	2018
► ASM Hell Learn basic assembly instructions through a game (LD41 submission).	2017
► Prolog KNN Implementation of kNN in Prolog, a language the least suited for this.	2017

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), visualization, 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