Project abbreviation: ELITR

Project name: European Live Translator

Project coordinator: Ondřej Bojar

Project consortium:

- Charles University
- University of Edinburgh
- Karlsruhe Institute of Technology
- PerVoice
- alfatraining

Project duration: January 1st, 2019 to March 31st, 2022.

Main key words: Machine translation, Multilingualism, language diversity, Automated translation, Natural language processing, Service oriented architectures

Background of the research topic: ELITR builds upon the experience of its partners in machine translation and speech processing and advances the state of the art by addressing the following challenges:

- Highly multi-lingual translation.
- Translation of spoken genre, live subtitling.
- Preservation of cross-sentence context, in written documents or spoken sessions.

Additionally, we are laying foundations for automatic meeting summarization ("minuting"), to provide the NLP community with one possible clear and concise definition and starting shared tasks on the topic, see [AutoMin 2021](#).

Goal of the project: Advancement and application of ASR and SLT in Face-to-face, Remote and Hybrid conferences, Robust and efficient systems for machine translation with many source and/or target languages, Coherent translation of larger documents and dialogues, Automatic minuting, i.e. summarizing meetings in the form of structured minutes from speech

Project abstract: [ELITR](#), an EU-funded research and innovation project pushes the frontiers of simultaneous speech translation and document-level translation.

Centered around use cases of two user partners, the Supreme Audit Office of the Czech Republic and alfatraining, ELITR is bringing to life:

- Live subtitling in up to 43 languages, with the major test event in April 2021: the [EUROSAI 2021](#) congress of supreme audit institutions,
- Speech translation in alfaview remote conferencing platform.

High quality natural language processing (NLP) is critically needed in the diverse European landscape.

ELITR is co-ordinated by [Charles University](#), with the [University of Edinburgh](#) and [Karlsruhe Institute of Technology](#) as two research partners, [PerVoice](#) as the integration partner and [alfatraining](#) as project-funded user partner. The [Supreme Audit Office](#) is contributing from its own funds.

Publications:

Ansari Ebrahim, Bojar Ondřej, Haddow Barry, Mahmoudi Mohammad: SLTev: Comprehensive Evaluation of Spoken Language Translation. In: Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations, Copyright © Association for Computational Linguistics (ACL), Stroudsburg, PA, USA, ISBN 978-1-954085-05-3, pp. 71-79, 2021

Bojar Ondřej: ELITR Coordination Experience. Talk, MUNI Grants Week, Brno, Czechia, Oct 2021

Bojar Ondřej: ELITR Demo at LCT Summer School. Talk, LCT Summer School 2021, Saarbruecken, Germany, Aug 2021

Bojar Ondřej: Strojové překlady a tlumočení. Talk, Seminář pro tlumočníky na ÚTRL, Praha, Czechia, Nov 2021

Bojar Ondřej, Macháček Dominik, Sagar Sangeet, Smrž Otakar, Kratochvíl Jonáš, Polák Peter, Ansari Ebrahim, Mahmoudi Mohammad, Kumar Rishu, Franceschini Dario, Canton Chiara, Simonini Ivan, Nguyen Thai-Son, Schneider Felix , Stüker Sebastian, Waibel Alex, Haddow Barry, Sennrich Rico, Williams Philip: ELITR Multilingual Live Subtitling: Demo and Strategy. In: Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations, Copyright © Association for Computational Linguistics (ACL), Stroudsburg, PA, USA, ISBN 978-1-954085-05-3, pp. 271-277, 2021

Bojar Ondřej, Srdečný Vojtěch, Kumar Rishu, Smrž Otakar, Schneider Felix , Haddow Barry, Williams Phil, Canton Chiara: Operating a Complex SLT System with Speakers and Human Interpreters. In: Proceedings of Machine Translation Summit XVIII 1st Workshop on Automatic Spoken Language Translation in Real-World Settings, Copyright © Association for Machine Translation in the Americas, Stroudsburg, PA, USA, pp. 23-34, 2021

Kopp Matyáš, Stankov Vladislav, Bojar Ondřej, Hladká Barbora, Straňák Pavel: ParCzech 3.0. Data/software, LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University, Prague, Czechia, http://hdl.handle.net/11234/1-3631, May 2021

Kopp Matyáš, Stankov Vladislav, Krůza Jan Oldřich, Straňák Pavel, Bojar Ondřej: ParCzech 3.0: A Large Czech Speech Corpus with Rich Metadata. In: 24th International Conference on Text, Speech and Dialogue, Copyright © Springer, Cham, Switzerland, ISBN 978-3-030-83526-2, pp. 293-304, 2021

Macháček Dominik, Žilinec Matúš, Bojar Ondřej: Lost in Interpreting: Speech Translation from Source or Interpreter?. In: Proceedings of INTERSPEECH 2021, Copyright © ISCA, Baxas, France, pp. 2376-2380, 2021

Bojar Ondřej: Promises and Pitfalls of Neural MT. Talk, Lunchtime session on Digital Transformation, Prague, Czechia,Nov 2018

Çano Erion, Bojar Ondřej: Efficiency Metrics for Data-Driven Models: A Text Summarization Case Study. In: Proceedings of the 12th International Conference on Natural Language Generation (INLG 2019), Copyright © Association for Computational Linguistics, Stroudsubrgh, PA, USA, ISBN 978-1-950737-94-9, pp. 229-239, 2019

Çano Erion, Bojar Ondřej: Keyphrase Generation: A Multi-Aspect Survey. In: Proceedings of the 25th Conference of Open Innovations Association FRUCT 2019, Copyright © Finnish-Russian University

[Click here to be redirected to the virtual room of the Project Expo](#).

Cooperation in Telecommunications, Helsinki, Finland, ISBN 978-952-69244-0-3, ISSN 2305-7254, pp. 85-94, 2019

Çano Erion, Bojar Ondřej: Keyphrase Generation: A Text Summarization Struggle. In: The 17th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Copyright © NAACL-HLT 2019, Minneapolis, USA, ISBN 978-1-950737-13-0, pp. 666-672, 2019

Kvapilíková Ivana, Macháček Dominik, Bojar Ondřej: CUNI Systems for the Unsupervised News Translation Task in WMT 2019. In: Fourth Conference on Machine Translation - Proceedings of the Conference, Copyright © Association for Computational Linguistics, Stroudsburg, PA, USA, ISBN 978-1-950737-27-7, pp. 241-248, 2019

Macháček Dominik, Kratochvíl Jonáš, Vojtěchová Tereza, Bojar Ondřej: A Speech Test Set of Practice Business Presentations with Additional Relevant Texts. In: Statistical Language and Speech Processing, Copyright © Springer Nature Switzerland AG, Cham, Switzerland, ISBN 978-3-030-31371-5, ISSN 0302-9743, pp. 151-161, 2019

Nedoluzhko Anna, Bojar Ondřej: Towards Automatic Minuting of Meetings. In: Proceedings of the 19th Conference ITAT 2019: Slovenskočeský NLP workshop (SloNLP 2019), Copyright © CreateSpace Independent Publishing Platform, Košice, Slovakia, ISBN 0000000000, ISSN 1613-0073, pp. 112-119, 2019

Pham Thuong-Hai, Macháček Dominik, Bojar Ondřej: Promoting the Knowledge of Source Syntax in Transformer NMT Is Not Needed. In: Computación y Sistemas, Vol. 23, No. 3, Copyright © Instituto Politécnico Nacional, ISSN 1405-5546, pp. 923-934, 2019

Popel Martin: Biases and perils of MT evaluation. Talk, Workshop on document level MT evaluation, Luxembourg, Luxembourg, Nov 2019

Rysová Kateřina, Rysová Magdaléna, Musil Tomáš, Poláková Lucie, Bojar Ondřej: A Test Suite and Manual Evaluation of Document-Level NMT at WMT19. In: Fourth Conference on Machine Translation - Proceedings of the Conference, Copyright © Association for Computational Linguistics, Stroudsburg, PA, USA, ISBN 978-1-950737-27-7, pp. 455-463, 2019

Vojtěchová Tereza, Novák Michal, Klouček Miloš, Bojar Ondřej: SAO WMT19 Test Suite: Machine Translation of Audit Reports. In: Fourth Conference on Machine Translation - Proceedings of the Conference, Copyright © Association for Computational Linguistics, Stroudsburg, PA, USA, ISBN 978-1-950737-27-7, pp. 680-692, 2019

Ansari Ebrahim, Axelrod Amittai, Bach Nguyen, Bojar Ondřej, Cattoni Roldano, Dalvi Fahim, Durrani Nadir, Federico Marcello, Federmann Christian, Gu Jiatao, Huang Fei, Knight Kevin, Ma Xutai, Nagesh Ajay, Negri Matteo, Niehues Jan, Pino Juan, Salesky Elizabeth, Shi Xing, Stüker Sebastian, Turchi Marco, Waibel Alex, Wang Changhan: FINDINGS OF THE IWSLT 2020 EVALUATION CAMPAIGN. In: Proceedings of the 17th International Conference on Spoken Language Translation, Copyright © Association for Computational Linguistics, Online, ISBN 978-1-952148-07-1, pp. 1-34, 2020

Barrault Loïc, Biesialska Magdalena, Bojar Ondřej, Costa-Jussà Marta R., Federmann Christian, Graham Yvette, Grundkiewicz Roman, Haddow Barry, Huck Matthias, Joanis Eric, Kocmi Tom, Koehn Philipp, Lo Chi-kiu, Ljubešić Nikola, Monz Christof, Morishita Makoto, Nagata Masaaki, Nakazawa Toshiaki, Pal Santanu, Post Matt, Zampieri Marcos: Findings of the 2020 Conference on Machine Translation (WMT20). In: Fifth Conference on Machine Translation - Proceedings of the Conference,

[Click here to be redirected to the virtual room of the Project Expo](#).

Copyright © Association for Computational Linguistics, Stroudsburg, PA, USA, ISBN 978-1-948087-81-0, pp. 1-55, 2020

Bojar Ondřej, Macháček Dominik, Sagar Sangeet, Smrž Otakar, Kratochvíl Jonáš, Ansari Ebrahim, Franceschini Dario, Canton Chiara, Simonini Ivan, Nguyen Thai-Son, Schneider Felix , Stüker Sebastian, Waibel Alex, Haddow Barry, Sennrich Rico, Williams Philip: ELITR: European Live Translator. In: Proceedings of the 22st Annual Conference of the European Association for Machine Translation (2020), Copyright © European Association for Machine Translation, Lisboa, Portugal, ISBN 978-989-33-0589-8, pp. 463-464, 2020

Çano Erion, Bojar Ondřej: Automating Text Naturalness Evaluation of NLG Systems.ArXiv.org Computing Research Repository, Copyright © Cornell University, ISSN 2331-8422, Jun 2020

Çano Erion, Bojar Ondřej: How Many Pages? Paper Length Prediction from the Metadata. In: 4th International Conference on Natural Language Processing and Information Retrieval, Copyright © ACM, New York, USA, ISBN 978-1-4503-7760-7, pp. 91-95, 2020

Çano Erion, Bojar Ondřej: Human or Machine: Automating Human Likeliness Evaluation of NLG Texts.ArXiv.org Computing Research Repository, Copyright © Cornell University, ISSN 2331-8422, Jun 2020

Çano Erion, Bojar Ondřej: Two Huge Title and Keyword Generation Corpora of Research Articles. In: Proceedings of the 12th International Conference on Language Resources and Evaluation (LREC 2020), Copyright © European Language Resources Association, Marseille, France, ISBN 979-10-95546-34-4, pp. 6663-6671, 2020

Franceschini Dario, Canton Chiara, Simonini Ivan, Schweinfurth Armin, Glott Adelheid, Stüker Sebastian, Nguyen Thai-Son, Schneider Felix , Ha Thanh-Le, Waibel Alex, Haddow Barry, Williams Phil, Sennrich Rico, Bojar Ondřej, Sagar Sangeet, Macháček Dominik, Smrž Otakar: Removing European Language Barriers with Innovative Machine Translation Technology. In: Proceedings of the 1st International Workshop on Language Technology Platforms, Copyright © ELRA, Paris, France, ISBN979-10-95546-64-1, pp. 44-49, 2020

Kratochvíl Jonáš, Polák Peter, Bojar Ondřej: Large Corpus of Czech Parliament Plenary Hearings. In: Proceedings of the 12th International Conference on Language Resources and Evaluation (LREC 2020), Copyright © European Language Resources Association, Marseille, France, ISBN 979-10-95546-34-4, pp. 6363-6367, 2020

Macháček Dominik, Bojar Ondřej: Presenting Simultaneous Translation in Limited Space. In: Proceedings of the 20th Conference Information Technologies - Applications and Theory (ITAT 2020), Copyright © Tomáš Horváth, Košice, Slovakia, ISSN 1613-0073, pp. 32-37, 2020

Macháček Dominik, Kratochvíl Jonáš, Sagar Sangeet, Žilinec Matúš, Bojar Ondřej, Nguyen Thai-Son, Schneider Felix , Williams Philip, Yao Yuekun: ELITR Non-Native Speech Translation at IWSLT 2020. In: Proceedings of the 17th International Conference on Spoken Language Translation, Copyright © Association for Computational Linguistics, Online, ISBN 978-1-952148-07-1, pp. 200-208, 2020

Polák Peter, Sagar Sangeet, Macháček Dominik, Bojar Ondřej: CUNI Neural ASR with Phoneme-Level Intermediate Step for Non-Native SLT at IWSLT 2020. In: Proceedings of the 17th International Conference on Spoken Language Translation, Copyright © Association for Computational Linguistics, Online, ISBN 978-1-952148-07-1, pp. 191-199, 2020

Rosa Rudolf: Cross-Lingual Information Retrieval. Data/software, ÚFAL MFF UK, Praha, Czechia,https://github.com/ufal/clir, Mar 2020

Rosa Rudolf: Cross-Lingual Information Retrieval DEMO. Data/software, ÚFAL MFF UK, http://bit.ly/ws-clir, Mar 2020

Rosa Rudolf: Deliverable D7.2 Report on NLP Technologies Workshop at EUROSAI Congress. Technical report no. 2020/D7.2, Copyright © ÚFAL MFF UK, Praha, Czechia, 25 pp., Jun 2020

Zouhar Vilém, Vojtěchová Tereza, Bojar Ondřej: WMT20 Document-Level Markable Error Exploration. In: Fifth Conference on Machine Translation - Proceedings of the Conference, Copyright © Association for Computational Linguistics, Stroudsburg, PA, USA, ISBN 978-1-948087-81-0, pp. 371-380, 2020