This document is part of the Research and Innovation Action "European Live Translator (ELITR)". This project has received funding from the European Union's Horizon 2020 Research and Innovation Programme under Grant Agreement No 825460.

Deliverable D1.2

Year 1 Test Sets

Philip Williams (UEDIN), Ondřej Bojar (CUNI), Jonáš Kratochvíl (CUNI), Dominik Macháček (CUNI), Anna Nedoluzhko (CUNI), Daniel Suchý (CUNI)

Dissemination Level: Public

Final (Version 1.0), 31st October, 2019

Grant agreement no.	825460
Project acronym	ELITR
Project full title	European Live Translator
Type of action	Research and Innovation Action
Coordinator	RNDr. Ondřej Bojar, PhD. (CUNI)
Start date, duration	1st January, 2019, 36 months
Dissemination level	Public
Contractual date of delivery	31st October, 2019
Actual date of delivery	31st October, 2019
Deliverable number	D1.2
Deliverable title	Year 1 Test Sets
Type	Report
Status and version	Final (Version 1.0)
Number of pages	9
Contributing partners	CUNI, UEDIN
WP leader	CUNI
Author(s)	Philip Williams (UEDIN), Ondřej Bojar (CUNI), Jonáš Kratochvíl (CUNI), Dominik Macháček (CUNI), Anna Nedoluzhko (CUNI), Daniel Suchý (CUNI)
EC project officer	Alexandru Ceausu
The partners in ELITR are:	▪ Univerzita Karlova (CUNI), Czech Republic ▪ University of Edinburgh (UEDIN), United Kingdom ▪ Karlsruher Institut für Technologie (KIT), Germany ▪ PerVoice SPA (PV), Italy ▪ alfatraining Bildungszentrum GmbH (AV), Germany
Partially-participating party	▪ Nejvyšší kontrolní úřad (SAO), Czech Republic

For copies of reports, updates on project activities and other ELITR-related information, contact:

RNDr. Ondřej Bojar, PhD., ÚFAL MFF UK
Malostranské náměstí 25
118 00 Praha, Czech Republic
bojar@ufal.mff.cuni.cz
Phone: +420 951 554 276
Fax: +420 257 223 293

Copies of reports and other material can also be accessed via the project's homepage:

`http://www.elitr.eu/`

Contents

1 Executive Summary

This deliverable reports on the ASR, MT, and Automatic Minuting test sets created during the first ten months of the project. Section 2 describes the ASR test sets, which take the form of audio speech data with reference transcripts:

- The Czech ASR tests sets (Section 2.1) complement the Czech ASR training data described in Deliverable 1.1. They are derived from recordings from Czech Radio and the European and Czech Parliaments.
- Antrecorp (Section 2.2) is a collection of recordings and transcripts of presentations given in English by non-native speakers from a range of EU countries. This test set is designed for testing the robustness of ASR systems to speaker accents and background noise.

The MT test sets are text sentences (or documents) paired with reference translations. They are described in Section 3:

- WMT19 Document-Level Test Suites (Section 3.1) assess the MT output quality for the domain of SAO.
- The News Commentary 14 Test Sets (Section 3.2) are out-of-domain test sets covering 63 language pairs.
- The TAUS In-Domain Test Sets (Section 3.3) cover eight language pairs. The data was selected according to its similarity to text in materials provided by participants to the WG VAT event.

Finally, Section 4 describes the Automatic Minuting test sets. These comprise audio recordings, transcriptions, and structured meeting minutes:

- The ELITR Meeting Minutes Test Sets (Section 4.1) is a collection of partially-annotated data gathered from project meetings held in English and Czech.

The Antrecorp test set and WMT19 Document-Level Test Suites have been publicly released and are available from the LINDAT/CLARIN repository and ELITR github page, respectively.

The News Commentary 14 Test Sets are sub-samples from publicly available data and will be made available in the Lindat repository and linked from ELITR web page (`http://lindat.cz` and `http://elitr.eu/`, respectively.)

The Czech ASR test sets and TAUS in-domain test sets are subject to copyright restrictions and can be used only internally within the project.

Meeting Minutes in the current form are suitable for internal evaluation and are user internally within the project. The annotation is ongoing and a public release will happen subsequently.

2 ASR Test Sets

2.1 Czech ASR Test Sets

As described in Deliverable 1.1, CUNI has been collecting and processing data for training Czech ASR systems. At the time of writing, they have processed 250 hours of audio and transcripts, with the majority coming from Český Rozhlas, the Czech radio broadcaster.

In order to test these ASR systems, CUNI has prepared three test sets. These include re-spoken plenary sessions from European parliament, recordings from Czech parliament meetings, and recordings from Český Rozhlas (see Table 1 for statistics, including the performance of the project's current Czech ASR system).

	Duration	WER
European parliament	2h 45m	6.37%
Czech parliament	2h 59m	6.70%
Český Rozhlas	2h 05m	9.21%

Table 1: Duration and model performance.

For the European Parliament test set, CUNI has also collected recordings of the same speeches in English, German, Spanish, and Italian. They plan to produce ASR transcripts from these and then to have them manually corrected, ideally by native speakers, to produce gold transcriptions. With test sets for the same speeches in multiple languages, the project will have a uniform test set for evaluating ASR systems and it will be possible to use the test set to evaluate SLT with, for instance, Czech audio as the input and English text as the output.

The Czech ASR test sets are subject to copyright restrictions and can be used only internally within the project. We are nevertheless searching for data that could be made publicly usable. Recordings from the Czech Parliament seem to be a good source and we plan to release a speech corpus based on them in the coming months.

2.2 Antrecorp

CUNI has created a corpus for evaluating ASR robustness based on short presentations given by high school students at the Fair of Student Firms, which took place in Prague in March 2019. All presentations were given in English by non-native speakers.[1] Recordings of the presentations were transcribed by the students after their speech and then later corrected to remove spelling errors, etc. The resulting corpus contains 39 recordings and transcriptions, along with related texts in the form of presentation slides and web pages provided by the participants. Each recording is between 40s and 120s and includes either one or two speakers. In total, 61 distinct speakers were recorded.

The test set has proven to be extremely challenging due to a combination of speaker accents and noise from the event (background music, people coming in and out of the presentation room, laughter, etc.). When tested on unadapted ASR systems, word error rates ranged from 20.9% to 100.0% (no output). Performance varied between systems, with the best results (mean 45.70% WER) being from a system trained on TED talks and Broadcast News using the Janus Recognition Toolkit.

The corpus has been made publicly available from the CLARIN/LINDAT repository[2] under the Creative Commons - Attribution 4.0 International license. A paper describing the work has been published at the International Conference on Statistical Language and Speech Processing in October 2019 [1].

[1]The first languages of the speakers were Czech, Swedish, Italian, German, Spanish, Romanian, Hungarian, Dutch, and Finnish.

3 MT Test Sets

3.1 WMT19 Document Level Test Suites

For the purposes of document-level evaluation, CUNI and SAO selected older audit reports published in several languages in the form of PDFs on the web sites of SAO and other supreme audit institutions.

CUNI and SAO cleaned this test suite and organized it in the form of suitable for WMT19 "test suites" task. The source texts were then distributed to all WMT19 News Task participants and thus translated by MT systems participating in the shared task. CUNI and SAO then assessed translation quality by the system using automatic and primarily manual evaluation.

The details about the test suite were published in the respective WMT19 paper [4].

The test suite itself has been made available as one of public ELITR repositories at github:

`https://github.com/ELITR/wmt19-elitr-testsuite`

3.2 News Commentary 14 Test Sets

As reported in Deliverable 1.1, UEDIN derived a multilingual corpus from OPUS to use for training the project's initial translation systems. The corpus covers the project's seven source languages[2] and the majority of the 42 target languages. In addition to training data, 2,000 sentence pairs of test data were reserved for each language pair. Unfortunately, these test sets were later found to contain a significant degree of cross-lingual overlap where the source (or target) sentence of a sentence pair in the training data for one language pair would occur as the source (or target) sentence in the test set of a different language pair. This is due to the presence in OPUS of a large number of sentences (usually originating in English) that have been multiply translated into several different languages. UEDIN plans to address this problem through cross-lingual filtering when preparing future versions of the multilingual corpus, but as a stopgap has created test sets extracted from the News Commentary v14 corpus. At the time of preparing the training corpus, version 14 of News Commentary was not included in the OPUS collection and so there is no overlap (apart from a few short, commonplace sentences) for the subset of News Commentary that was newly added to version 14.

The News Commentary corpus includes document boundaries, with documents typically being 30-50 sentences long. UEDIN prepared the test sets by randomly sampling documents until there were at least 2,000 sentence pairs for each language pair. While sampling they checked for overlap with the OPUS training corpus, only selecting documents with no overlapping sentences.

Since the News Commentary corpus does not cover all of the language pairs of interest to the project, it was only possible to produce test sets for a subset of the 287 language pairs, but this was sufficient for testing of the project's preliminary multilingual systems. For each of the seven source languages, there are test sets for Arabic, Czech, Dutch, English, French, German, Italian, Portuguese, Russian, and Spanish, making a total of 63 test sets.

As already noted, these test sets are only intended for temporary use, at least for primary testing. Once more suitable in-domain test sets have been created, they will be used as secondary test sets (where they can be used to test that we are not overfitting too severely to the target domain).

The test sets have only been shared internally.[3] The corpus will be made available in the Lindat repository and linked from ELITR web page (`http://lindat.cz` and `http://elitr.eu/`, respectively).

[2] The source languages are English, Czech, German, French, Italian, Spanish, and Russian.

[3] `http://data.statmt.org/elitr/nc14-test/v1/` The website is password protected—contact the authors of this document for access.

3.3 TAUS In-Domain Test Sets

Ahead of the VAT working group event in June 2019, CUNI collected slides and other relevant materials from participants and used these documents to search for similar sentences in the TAUS repository. From this data, they created an in-domain parallel corpus, covering eight language pairs: German to Czech as well as seven pairs with English as the source language and Czech, German, Spanish, Hungarian, Dutch, Polish, and Romanian as the target language. Dev and test sets of 2000 sentence pairs were randomly selected, with the rest being available for use as training data.

There was insufficient time to use this data to adapt our models for the June event, but CUNI subsequently used the English-Czech portion for preliminary domain adaptation experiments. Using the training portion of the in-domain TAUS data, they fine-tuned out-of-domain models that were trained on the OPUS corpus described in Deliverable 1.1, seeing promising improvements in translation quality when evaluating on the in-domain test sets.[4]

Due to licensing restrictions, this data has only been shared internally.[5]

[4]For a bilingual English-Czech system, the BLEU score on the TAUS in-domain set increased from 24.8 to between 27.1 and 34.3, depending on the fine-tuning strategy that was used. As expected, this came at the cost of a decrease in performance on the out-of-domain News Commentary test set, which fell from 29.1 to between 10.4 and 22.3, again depending on the fine-tuning regime. Future work will investigate strategies that reduce the loss in general domain translation quality, while optimising in-domain performance.

[5]`https://drive.google.com/drive/folders/1H2vrkdfm8jNO99q9Y5CCb7zdPA-SujnM`

4 Automatic Minuting Test Sets

4.1 ELITR Meeting Minutes Test Sets

During the first ten months of the ELITR project duration, CUNI started collecting data for the corpus of Meetings and Minutes. The data includes the audio recordings, ASR transcripts, and pre-prepared agendas, together with meeting minutes created by the meeting organizers or a secretary after the meeting.

So far, approximately 40 hours of meetings in English and 6 hours of meetings in Czech have been collected (see Table 2). The participants of meetings in English are mostly non-native speakers, thus the data represent a special challenge for ASR.

The obtained data are further manually annotated by two annotators. For the annotation, CUNI used the NITE XML Toolkit, which was used for the AMI meetings annotation [3]. The tool has been slightly improved to satisfy the needs of the project. The main idea of the annotation is to connect segments in the audio recordings and ASR transcripts with the corresponding items in the agenda. So far 10 meetings have been annotated (approx. 12 hours of recordings); one meeting has been annotated by two annotators in order to be able to assess inter-annotator agreement.

The corpus is still under development. At the time being, it has only been used for our preliminary experiments in meeting summarization. Due to the small amount of annotated data so far, the whole corpus was used as a test set for the experiments.

	Number	Duration	Minutes	Annotations
ELITR internal meetings	27	34h 18m	yes	partly
BERGAMOT internal meetings	4	5h 47m	yes	not yet
LSD and other in Czech	5	5h 53m	yes	not yet

Table 2: ELITR Meetings and Minutes Corpus Statistics (in progress)

The corpus will be divided into train, development and test sections and publicly released under a permissive license as soon as a reasonable size for training is reached and the data will be checked to conform to GDPR requirements.

References

[1] Dominik Macháček, Jonáš Kratochvíl, Tereza Vojtěchová, and Ondřej Bojar. A speech test set of practice business presentations with additional relevant texts. In Carlos Martín-Vide, Matthew Purver, and Senja Pollak, editors, *Statistical Language and Speech Processing*, pages 151–161, Cham, 2019. Springer International Publishing. ISBN 978-3-030-31372-2.

[2] Dominik Macháček, Jonáš Kratochvíl, Tereza Vojtěchová, and Ondřej Bojar. A speech test set of practice business presentations with additional relevant texts, 2019. URL `http://hdl.handle.net/11234/1-3023`. LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

[3] I. Mccowan, G. Lathoud, M. Lincoln, A. Lisowska, W. Post, D. Reidsma, and P. Wellner. The ami meeting corpus. In *In: Proceedings Measuring Behavior 2005, 5th International Conference on Methods and Techniques in Behavioral Research. L.P.J.J. Noldus, F. Grieco, L.W.S. Loijens and P.H. Zimmerman (Eds.), Wageningen: Noldus Information Technology*, 2005.

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