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 D1.4

Year 2 Test Sets

Daniel Suchý, Anja Nedoluzhko, Ondřej Bojar (CUNI)

Dissemination Level: Public

Final (Version 1.0), 4th September, 2020

Grant agreement no.	825460
Project acronym	ELITR
Project full title	European Live Translator
Type of action	Research and Innovation Action
Coordinator	Doc. RNDr. Ondřej Bojar, PhD. (CUNI)
Start date, duration	1st January, 2019, 36 months
Dissemination level	Public
Contractual date of delivery	31st August, 2020
Actual date of delivery	4th September, 2020
Deliverable number	D1.4
Deliverable title	Year 2 Test Sets
Type	ORDP: Open Research Data Pilot
Status and version	Final (Version 1.0)
Number of pages	10
Contributing partners	CUNI, UEDIN
WP leader	UEDIN
Author(s)	Daniel Suchý, Anja Nedoluzhko, Ondřej Bojar (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:

Doc. 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 test sets that have been created and are in use at month 20 of the project. Since Deliverable 1.2 (Year 1 Test Sets), a major focus of our effort has been to compile our ASR, MT, and SLT test sets into a single high-quality collection with a clearly defined and consistent structure. The result is `elitr-testset`, a public GitHub repository.[1] Section 2.1 provides an overview of the individual test sets currently contained in the `elitr-testset` repository and Section 2.2 describes our processes of quality control along with some benchmark evaluation results.

It is important to mention that `elitr-testset` is meant to be a gradually growing collection, versioned essentially by commit IDs, so that it can be continuously expanded with more languages and more files while allowing for a full reproducibility of test results performed with this dataset.

Additionally, we report on the progress of data acquisition for WP5 Minuting. The first report on this dataset was in December 2019 (see *D1.3: Internal Release of Corpus of Minutes*) and we now have considerably more data, see Section 3. The minuting data however cannot be publicly released yet because it still has to be anonymized.

[1] `https://github.com/ELITR/elitr-testset`

2 The ELITR Test Set Repository

Given the complexity of the main intended ELITR use case, the testing alone of the system components is challenging task. Ideally, ELITR should be translating speech from 7 source into up to 43 target languages in one domain and further domains are needed for additional test situations which we are adding on the go. Therefore, we are not creating a single test set, but rather a collection of inputs and reference translations and transcripts with an agreed structure.

We call this collection `elitr-testset` and keep it versioned in a public GitHub repository: `https://github.com/ELITR/elitr-testset`

We rely on git versioning, so individual commits, identified by git commit IDs, are versions of the test set. Any addition or modification of the sources or references is thus clearly preserved and an exactly comparable re-evaluation of a system remains always possible as long as the commit ID of the test set is known.

`elitr-testset` also supports "subsetting", i.e. testing only on agreed subsets of materials. This subsetting will be used particularly for domain-specific evaluation. More details on this technique will be provided in the README of the repository, in the section on "indices".

We are also in the process of linking `elitr-testset` with our development and integration pipelines, so that our systems are going to be regularly tested in terms of ASR and MT quality.

For public use, we plan to directly link our SLT evaluation tool, SLTev[2] with `elitr-testset`, so that fellow researchers can easily see how their system performs on our data.

In the rest of this section, we describe the data composition (Section 2.1) and the quality control mechanisms we apply to the collection (Section 2.2).

2.1 Description of Test Set Data

Here, we provide a brief overview of all the subsets of our current test set data. As `elitr-testset` will grow, further subsets will be included.

Some of the subsets are grouped together because the origin of the data was the same, some are grouped because they were added for a common purpose, and other reasons behind new groups will emerge in the future. The layout of `elitr-testset` is designed to allow this flexibility in adding and using any subset of the data.

2.1.1 Czech ASR

This Czech monolingual dataset contains 5 separate recordings. Four of the recordings were acquired from publicly available sources, such as the Czech and European parliaments, or the Czech radio. The material for the fifth comes from the WG-VAT (Working Group Value Added Tax) conference held in 2019 by the user partner EUROSAI. Four recordings are prepared speeches on different domains: One on auditing and three parliamentary speeches. The last one is a presentation on machine translation. The dataset contains over 10 hours of audio. It includes manual transcriptions by native speakers. This data set is finished and ready to be used for evaluating our systems.

2.1.2 Less-Resourced Languages

This dataset[3] consists of texts in languages that are part of the ELITR scope but were not handled in the ELITR project so far. It includes texts that were not originally in English, Czech or German. In some cases, the texts are multilingual documents, published in many languages. In other cases, the there is an original text with single or multiple translations. All are at least bilingual, so that we do not need to spend resources on translation, although this last resort option may become necessary in some cases. All sources are publicly available, and

[2] `https://github.com/ELITR/SLTev`

[3] Interally and colloquially, we sometimes refer to this dataset as "exotic languages", although there is obviously nothing exotic about them.

were acquired from the websites of various national and European institutions, such as Court of justice of the European union, Russian ministry of foreign affairs or Barcelona city council. The data is processed to follow the same format and then verified by a native speaker of the language. All texts in a set are verified to be parallel with each other. The data includes written reports as well as transcripts of public speeches in the domain of auditing or government. No audio is included. This data set is a work in progress, more relevant sources are still being searched for across the under-resourced languages and the identified data are still being processed. The table below describes the sentence counts for processed sources:

language	sentence count
Bosnian	230
Danish	1014
Dutch	1014
Catalan	842
English	2920
Serbian	834
Spanish	842

The table shows how many sentences we have in our processed, parallel files. All documents in a set are parallel to each other. The sources aren't English-centric, meaning that we have e.g. danish-dutch parallel texts.

2.1.3 Non-native Speech Translation at IWSLT 2020

This is the largest and the most polished of our data sets, originally used as a development and test set for the IWSLT shared task on non-native speech translation, which took place in July 2020. The data set consists of 5 parts with different sources which are described in detail in the appendix of Ansari et al. (2020).

- Antrecorp – data from collected at the Fair of Student Firms, which took place in Prague in March 2019. It contains 1 hour and 18 minutes of audio, transcribed into 571 sentences.
- Khan Academy – publicly available at https://www.khanacademy.org/. It contains 25 minutes of audio, transcribed into 538 sentences.
- `sao-consecutive` – consecutive interpretation data made publicly available by the Supreme audit office of the Czech Republic. It contains 21 minutes of audio, transcribed into 200 sentences.
- `sao-wgvat` – data from the WG-VAT meeting held in 2019. Permission to use this data for IWSLT and further was given by the Supreme Audit Office of Czech Republic and the respective speakers. It contains 1 hour and 15 minutes of audio, transcribed into 455 sentences.
- AMI corpus (Mccowan et al., 2005)[4] – data made publicly available as a part of the AMI corpus. Our subselection contains 1 hour and 25 minutes of audio, transcribed into 1516 sentences.

All the sources were processed to follow the same format and the same guidelines. English is the source language for all data in this dataset, with translations into Czech and German. In case of SAO consecutive interpretation, SAO WG-VAT and Khan Academy, some Czech translations were already a part of the source data. The type of translation differed across these subsets: The WG-VAT data included simultaneous interpretation (incl. manually revised transcripts) by students of interpretation provided by ELITR from our first test event. For the

[4] `http://groups.inf.ed.ac.uk/ami/corpus/`

purposes of IWSLT, we complemented it with professional translation of the English transcripts. The other SAO set includes transcripts of consecutive interpretation into Czech but again, we equipped it with another professional translation of the English transcript. The Khan Academy subset consists of textual translations. Czech translations for Antrecorp and AMI were created by the ELITR project, German translations for all the sources of data were made by KIT.

All sources of data include the original audio or audio / video source, in the original format. The original audio was manually transcribed. Then, force-alignment was created for the files. In some cases the automatic forced alignment failed. We manually segmented these files into word-length segments.

The data consists mostly of spoken, prepared presentations, except for the AMI corpus, which contains spontaneous discussions between the participants of the AMI project. The SAO data is from the auditing domain, the other sources are from other various domains.

This dataset is finished and ready for to be used for evaluating our systems.

2.1.4 LangTools Workshop Dry-Run Sessions

These are the dry-run sessions of "LangTools" workshop, a workshop prepared by CUNI for the EUROSAI Congress in Prague. The workshop is described in detail in the deliverable *D7.2: Report on NLP technologies Workshop at EUROSAI Congress*. Due to Covid-19, the congress was postponed to 2021 and the workshop did not take place yet, but two dry runs did.

The data consists of several spoken presentations in English, where researchers from UFAL present the purpose of the ELITR project, or discuss various issues related to machine translation. This dataset is a work in progress, and we are currently working on processing the audio, transcribing the speeches and then manually segmenting the recordings. We would also like to create several translations for this dataset, including translations into languages which were not broadly represented in the ELITR test set so far. Currently, five speeches are processed and transcribed, totalling 47 minutes of audio and 314 sentences.

2.1.5 Linguistic Mondays

This dataset consists of a single recording of a presentation given at the Linguistic Mondays seminars, a workshop held at UFAL. It contains a spoken presentation in English, in the domain of machine translation. Total audio time is 1 hour and 17 minutes, and it was manually transcribed into 846 sentences. An original video recording is also available at the UFAL website. The work on this data set is finished, but it can be expanded by acquiring additional translations or additional recordings from the Linguistic Mondays seminar.

2.1.6 TAUS

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.[5] From this data, we 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. Development and test sets of 2000 sentence pairs each were randomly selected, with the rest being available for use as training data. This dataset contains 4000 sentences for each language pair, totalling 3200 sentences. The data is extracted from documents discussing government and finance.

2.1.7 WMT 19 ELITR Test Suite

These are the files that we sent to the Fourth conference on machine translation (WMT19). The data consists of 11 written reports (4741 sentences) in the domain of auditing. Five of

[5] `https://www.taus.net/`

the documents have English as the original language, the others are in Czech. The English documents were professionally translated into Czech and vice versa. All of the documents were also translated into German. This data set is finished and ready for use.

2.2 Evaluation of Test Set data

2.2.1 Automatic Quality Control

As a part of gathering and processing the data, we have set up a system of automatic quality control checks. These scripts verify data integrity and format of all the data included in the `elitr-testset` repository. This enables us to avoid human mistakes and maintain a structure and high quality of data in an ever-expanding dataset. We have checked for the following types of errors:

- file scheme – all files follow the same naming scheme. Only certain file types (extensions) are allowed in the repository
- textual integrity – all translated files come with the original source and are parallel with this source. Translation is complete, e.g. without blank sentences. Files are in a one-sentence-per-line format.
- technical integrity – text files are encoded either with ASCII or UTF-8 encodings, and have the same (UNIX) line endings. Audio files are compressed, valid, and under a specified size limit to prevent the repository from growing excessively.[6]

Many errors have been found and corrected thanks to these checks. There are some minor inconsistencies remaining, such as imperfect sentence alignment (multiple sentences per line) in some of the text files, or inconsistent headers in audio files.

2.2.2 Evaluation of ASR and MT

To illustrate the applicability of `elitr-test set` for our purposes, we provide sample evaluations.

The following table summarizes the results of our ASR and MT systems on various parts of our test set. As the test set will grow in language and domain coverage and as we will be obtaining new versions of our systems, the results will be expanded and will allow for comparison. For now, the scores can only serve as baseline, not directly comparable to each other. The aggregate results are computed over all relevant files in the corpus, e.g. 'avgWER' summarizes the results of our ASR systems for all recordings in the particular data set. BLEU scores are measured for MT files only. This table disregards that some files have multiple references and aggregates results of both target languages (cs/de), and thus the scores can serve only as a sanity check and overall regression test. In absolute terms, both WER and BLEU scores are rather bad, mainly due to the strong accent of English non-native speakers and domain differences for translation. For completeness, `elitr-testset` was at commit 5bc5a4e577ad648a5a453cfc0edf60b4ad3a18ff and the evaluation tool SLTev was at commit f1017fc5ae5547492a570a33d0f04fd63a0fff75 when we ran this evaluation.

data set	source lang	target lang	avgBLEU	stddevBLEU	avgWER	stddevWER
iwslt-devset	en	cs/de	9.33	4.13	83.05	3.01
iwslt-antrecorp	en	cs/de	8.02	4.07	78.49	14.21
iwslt-consecutive	en	cs/de	6.49	5.21	91.81	5.46
iwslt-khanacademy	en	cs/de	5.00	1.93	60.31	5.59
iwslt-wgvat	en	cs/de	12.36	3.36	85.27	16.12

[6]Files over 50MB are kept outside of the repository and the repository only contains a URL or other link to them.

3 Test Data for Automatic Minuting

We are creating the corpus of meetings and meeting minutes which for the purposes of experiments on automatic minuting. We collect internal and partly external data. Internal meetings include the ELITR internal meetings, as well as other meetings which take part within the ELITR partners community (mostly be the CUNI colleagues). External meetings are collected by the Prague Institute of Planning and Development.

The corpus consists of audio or video recording of meetings, automatic transcripts (ASRs), manually checked and corrected transcripts and the minutes for these meetings. The minutes include original minutes of the meeting (created by the meeting leader or the secretary) and manually generated minutes by the annotators. We also collect several variants of minutes for the same meetings, in order to be able to measure the similarity between them and to distinct them automatically from other minutes for different meetings.

So far, approximately 101 hours of meetings in English and 54 hours of meetings in Czech have been collected. The participants of meetings in English are mostly non-native speakers, thus the data represents a special challenge for ASR.

The annotation of the minuting corpus is going on and in will be further increased. The corpus statistics for the time being is presented in the table below.

	Meetings	Duration	Manually transcribed	Minutes
English	82	101 h	69 meetings / 72 h	63 meetings / 72 h
Czech	48	54 h	46 meetings / 50 h	46 meetings / 50 h

References

Ebrahim Ansari, Amittai Axelrod, Nguyen Bach, Ondřej Bojar, Roldano Cattoni, Fahim Dalvi, Nadir Durrani, Marcello Federico, Christian Federmann, Jiatao Gu, Fei Huang, Kevin Knight, Xutai Ma, Ajay Nagesh, Matteo Negri, Jan Niehues, Juan Pino, Elizabeth Salesky, Xing Shi, Sebastian Stüker, Marco Turchi, Alexander Waibel, and Changhan Wang. FINDINGS OF THE IWSLT 2020 EVALUATION CAMPAIGN. In *Proceedings of the 17th International Conference on Spoken Language Translation*, pages 1–34, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.iwslt-1.1. URL `https://www.aclweb.org/anthology/2020.iwslt-1.1`.

Iain Mccowan, J Carletta, Wessel Kraaij, Simone Ashby, S Bourban, M Flynn, M Guillemot, Thomas Hain, J Kadlec, Vasilis Karaiskos, M Kronenthal, Guillaume Lathoud, Mike Lincoln, Agnes Lisowska Masson, Wilfried Post, Dennis Reidsma, and P Wellner. The ami meeting corpus. *Int'l. Conf. on Methods and Techniques in Behavioral Research*, 01 2005.