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 D1.1

Initial Training Data, Separating Confidential and Public Version.

Jonáš Kratochvíl (CUNI), Barry Haddow (UEDIN), Philip Williams (UEDIN)

Dissemination Level: Public

Final (Version 1.0), 30th June, 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	30th June, 2019
Actual date of delivery	30th June, 2019
Deliverable number	D1.1
Deliverable title	Initial Training Data, Separating Confidential and Public Version.
Type	Report
Status and version	Final (Version 1.0)
Number of pages	9
Contributing partners	CUNI, UEDIN
WP leader	CUNI
Author(s)	Jonáš Kratochvíl (CUNI), Barry Haddow (UEDIN), Philip Williams (UEDIN)
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-parcitipating 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 public and internal training corpora created during the first six months of the project. These corpora comprise ASR and MT data collected from multiple external sources and then curated and processed ready for use within the project. The corpora described here were produced as part of Tasks 1.1 and 1.3:

- Our collection of Czech ASR data is described in Section 2. It is a curated version of a pre-existing speech training data comprising approximately 525 hours of Czech-language audio and corresponding transcriptions. About 75 hours of this data has been processed into a common format ready as part of Task 1.1 and successfully used in the training of an initial Czech ASR model. We will continue the work to further improve the ASR quality.
- The ELITR SAI Crawled Corpus v1.0 (Section 3) is collection of in-domain monolingual text corpora for 26 EUROSAI languages. The data was crawled from EUROSAI websites using the ParaCrawl pipeline. This corpus was created during phase one of Task 1.3, which ran in the months 1–6. A second phase of data collection is scheduled to run in the months 19–24 and so the existing corpus will be refined and extended according to the results of initial evaluation and the needs of the project at that time.
- The ELITR OPUS Corpus v1.0 (Section 4) is collection of out-of-domain parallel text corpora covering the project's seven source languages and 43 target languages. The corpora were sampled from the OPUS collection, with preference given to the subcorpora that more closely match the domain of public administration. As with the crawled corpus, the current OPUS corpus is a first revision, with a second phase of data collection scheduled to run from months 19–24.

The speech data are subject to copyright restrictions and can be used only internally within the project.

The text data (ELITR SAI Crawled Corpus v1.0 and ELITR OPUS Corpus v1.0) are subsamples 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/`, resp.)

2 Czech ASR Data

Ultimately, ELITR will support automatic speech recognition for seven languages: English, German, French, Spanish, Italian, Russian, and Czech. At the outset of the project, Czech had the weakest existing ASR support within the consortium and required the development of a new speech recognition pipeline as well as collection of the training data. Due to the high cost of speech transcription, it is not feasible to create new ASR training data within the project (beyond the small-scale data used for adaptation or evaluation) and so the goal of Task 1.1 was to gather pre-existing speech training data and process it to a common format ready to train the acoustic models.

The main source of acoustic data for Czech language are the recordings obtained from Český rozhlas (Czech Radio), the Czech national broadcast station. These recordings come from various distinct radio programs ranging from morning news to political debates. The estimated total length of audio samples is 200 hours of speech. All recordings have also corresponding text transcriptions, which are available in very high quality as Český rozhlas appoints a specialized company to transcribe them. The speech quality is also fairly good as the audio is recorded in very sterile environment without any background noise and with the usage of professional sound equipment.

Unfortunately each of the provided recordings are between 20-40 minutes long, which is unfeasible length for acoustic model training. Due to this fact a whole new force-alignment pipeline had to be developed to segment the original recordings into shorter ones (10-15 seconds in length) together with the corresponding transcriptions so that they can be used for ASR training.

These data preparation steps are still in progress. As of now, we have 58 000 audio files, each of which corresponds to one utterance, and the corresponding transcription. The total time of these processed data is 73 hours and 30 minutes. We have already successfully tested its applicability by training our Czech ASR model for the first ELITR event: WG VAT Workshop at SAO which took place on June 27 and 28, 2019.

Together with the data from Český rozhlas, CUNI is planning to use the Czech version of Vystadial corpus [1], which was developed specifically for ASR training and contains 15 hours of transcribed telephone conversations.

3 The ELITR SAI Crawled Corpus v1.0

The ELITR project is primarily aimed at the domain of public administration. Ideally, our MT and SLT systems would all be trained from in-domain parallel data, but such data is not readily available. In order to supplement our mixed domain training corpus, we have collected in-domain data from all the websites of all the supreme audit organisations in EUROSAI, to give us a monolingual in-domain corpus in all languages of interest. In-domain monolingual data is useful for MT training, since it can be used to construct synthetic parallel data, and then combined with natural parallel data. Alternatively, the monolingual data can serve as a sample of the type of text we wish to translate, and then can be used to select training data using Information Retrieval techniques.

3.1 Methodology

The corpus collection used a version of the ParaCrawl[1] crawling pipeline bitextor.[2] The bitextor pipeline is actually for extracting parallel corpora from web crawls, and contains all the components from the identification of promising sites, through crawling, extraction, alignment and cleaning. For the purposes of creating this corpora, we only needed crawling, extraction, language identification and some basic cleaning. We used the list of EUROSAI members[3] to identify sites for crawling, and also crawled EUROSAI itself. We crawled PDF files as well as HTML, but so far the corpus only includes text extracted from HTML since, at the time of writing, PDF extraction had only just been added to bitextor. We extracted text for all languages supported by the Moses[4] sentence splitting tools.

3.2 Corpus Statistics and Availability

In the table below we show the sizes of the corpora in terms of sentence pairs, broken down by language.

So far, the prepared corpus has only been shared internally.[5] The corpus will be made available in the Lindat repository and linked from ELITR web page (`http://lindat.cz` and `http://elitr.eu/`, respectively).

[1] `http://www.paracrawl.eu`
[2] `https://github.com/bitextor/bitextor`
[3] `https://www.eurosai.org/en/about-us/members/`
[4] `http://www.statmt.org/moses`
[5] `http://data.statmt.org/elitr/sai-crawls/monolingual/v1/` The website is password protected - contact the authors of this document for access.

Language	Cleaned	Deduped
Bulgarian	4507	3089
Bosnian	1980	626
Czech	304807	32048
Danish	7544	4871
Greek	1280	870
English	1122063	192075
Estonian	209369	41564
French	204546	21399
Croatian	1833	1306
Icelandic	10948	9661
Italian	741440	83035
Lithuanian	64848	22484
Dutch	202189	2468
Polish	1016196	102068
Portugese	227504	13523
Romanian	117059	42056
Russian	437529	114938
Slovakian	1288528	6760
Slovenian	40087	14841
Albanian	37031	7078
Serbian	317984	8630
Swedish	18793	15229
Turkish	7558	3317
Ukrainian	738765	218907

4 The ELITR OPUS Corpus v1.0

ELITR will provide translation support for seven source languages and up to 42 target languages.[6] While in-domain parallel data may be unavailable for most language pairs, the OPUS collectionn[7] provides a rich source of out-of-domain (and often 'similar-domain') parallel data. In this initial phase of data collection, we have aimed for broad coverage, with our initial parallel training corpus covering many language directions and genres. A subset of the resulting corpus was used to train the multilingual system deployed for the VAT Working Group meeting in June 2019.

4.1 Methodology

Of the 287 language pairs[8] of interest, the OPUS collection currently provides parallel data for 266. We used OPUS's web-based API to query the collection database and download all of this data.

For each language pair, OPUS data is divided between corpora covering a wide variety of domains. We began by splitting each domain-specific corpus into training, development, and test sections, using 2,000 randomly chosen sentence pairs each for development and test and leaving the rest for training. If a language pair had less than 16,000 sentence pairs in total, we did not split the corpus (to allow at least a 80% / 20% ratio between train and dev/test), instead taking all of the data as training data.

Of the available domains, some are closer than others to the target domain of public administration. There are also large data size imbalances between domains and language pairs. For

[6] For the purposes of this corpus, we do not differentiate between Moldovan and Romanian languages.
[7] `http://opus.nlpl.eu`
[8] 287 = 7 x 41, since all seven source languages are also target languages.

instance, for English-German there are 36,466,571 sentence pairs of OpenSubtitles but about a tenth of the volume – 3,264,676 – for EuroParl. But English-German EuroParl swamps the combined Belarusian-Italian data, which comprises 80,644 sentence pairs across all domains.

Since the principal application for this corpus is to build our baseline multilingual machine translation systems, we downsampled the data from the most highly-resourced languages so that the total data size across all languages would be of a manageable size and the languages would be balanced. We sampled up to 1 million sentence pairs per language pair, drawing samples first from Europarl (if available), then EUbooks, OpenSubtitles, and finally all remaining domains. This ordering was chosen based on the similarity to the target domain (taking into account the use case of spoken language translation). For language pairs with more than 100,000 sentence pairs, but less than a million, we upsampled the data to obtain exactly one million pairs. In this first version of the training data, we did not include language pairs where there were less than 100,000 pairs.

4.2 Corpus Statistics and Availability

The resulting corpus contains 226M sentence pairs of training data (1M for each pair that is covered). The following table shows for each of the seven source languages, how many pairings are present for the 24 EU languages, and similarly for the 18 non-EU EUROSAI languages). The numbers of currently unsupported pairings are given in parentheses.

Source	EU	EUROSAI	Not Covered
English	23 (0)	13 (5)	Belarusian, Armenian, Kazakh, Luxembourgish, Montenegrin
German	23 (0)	13 (5)	Belarusian, Armenian, Kazakh, Luxembourgish, Montenegrin
Czech	23 (0)	12 (6)	Azerbaijani, Belarusian, Armenian, Kazakh, Luxembourgish, Montenegrin
French	23 (0)	12 (6)	Azerbaijani, Belarusian, Armenian, Kazakh, Luxembourgish, Montenegrin
Spanish	23 (0)	12 (6)	Azerbaijani, Belarusian, Armenian, Kazakh, Luxembourgish, Montenegrin
Italian	23 (0)	12 (6)	Azerbaijani, Belarusian, Armenian, Kazakh, Luxembourgish, Montenegrin
Russian	23 (1)	12 (5)	Belarusian, Armenian, Kazakh, Luxembourgish, Maltese, Montenegrin

So far, the prepared corpus has only been shared internally.[9] The corpus will be made available in the Lindat repository and linked from ELITR web page (`http://lindat.cz` and `http://elitr.eu/`, respectively).

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

References

[1] Matěj Korvas, Ondřej Plátek, Ondřej Dušek, Lukáš Žilka, and Filip Jurčíček. Free English and Czech telephone speech corpus shared under the CC-BY-SA 3.0 license. In *Proceedings of the Eigth International Conference on Language Resources and Evaluation (LREC 2014)*, page To Appear, 2014.