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 D1.5

Final Training Data, Separating Confidential and Public Version

Philip Williams (UEDIN), Barry Haddow (UEDIN), Rico Sennrich (UEDIN), Dominik Macháček (CUNI), Anna Nedoluzhko (CUNI), Jonáš Kratochvíl (CUNI), Thai-Son Nguyen (KIT), Rishu Kumar (CUNI), Daniel Suchý (CUNI), Tirthankar Ghosal (CUNI), Ondřej Bojar (CUNI)

Dissemination Level: Public

Final (Version 1.0), 31st December, 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 December, 2020
Actual date of delivery	31st December, 2020
Deliverable number	D1.5
Deliverable title	Final Training Data, Separating Confidential and Public Version
Type	ORDP: Open Research Data Pilot
Status and version	Final (Version 1.0)
Number of pages	24
Contributing partners	CUNI, UEDIN, KIT
WP leader	UEDIN
Author(s)	Philip Williams (UEDIN), Barry Haddow (UEDIN), Rico Sennrich (UEDIN), Dominik Macháček (CUNI), Anna Nedoluzhko (CUNI), Jonáš Kratochvíl (CUNI), Thai-Son Nguyen (KIT), Rishu Kumar (CUNI), Daniel Suchý (CUNI), Tirthankar Ghosal (CUNI), 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 public and internal training corpora created during the first two years of the project. These corpora comprise ASR, MT, SLT, and AM data collected from multiple external sources, then curated and processed and now ready for use within the project:

ASR The project's ASR data collection efforts have focused on the gathering and processing of transcribed Czech speech data to create new corpora and on the compilation of existing English corpora into a multi-domain corpus suitable for testing domain robustness. Details are given in Section 2.

MT In order to provide MT training data for the research and integration systems required by the project, we concentrate on two areas: achieving high language pair coverage by curating large scale multi-domain multilingual datasets and collecting new in-domain data corpora through targeted web crawling and sentence alignment. See Section 3 for details.

SLT Compared to ASR and MT, datasets for SLT are scarce. One under-utilized source of data is the plenary sessions of the European Parliament. We have processed and aligned large volumes of speech data with their translations for all EU languages. Details are given in Section 4.

AM For AM, we have been collecting recordings and minutes from project meetings (including internal ELITR meetings) to form the ELITR Corpus of Minuting. In addition to self-collected data, we have compiled selected European Parliament Data in English paired with their corresponding minutes. See Section 5 for details.

In addition to the training data above, we have continued to update the `elitr-test-set` collection. We give details in Section 6.

2 ASR Training Data

For ASR, the project's data collection efforts have focused on two areas: i) the gathering and processing of transcribed Czech speech data in order to create new corpora; and ii) the compilation of existing English corpora into a multi-domain corpus suitable for testing domain robustness.

2.1 Czech Data

At the outset of the project, publicly available training corpora for the Czech language were scarce and there was a lack of pre-existing data within the consortium. Due to the high cost of transcription, it was not feasible to commission annotation work (beyond the small scale used for adaptation or evaluation) and so our goal was to gather pre-existing audio with 'naturally-occurring' transcripts and process them to a common format ready for train acoustic models.

CUNI began with recordings obtained from Český rozhlas, the Czech national radio broadcast station. These recordings come from various distinct radio programs ranging from morning news to political debates. All recordings have corresponding text transcriptions which are of very high quality as Český rozhlas appoints a specialised company to create the transcription. The speech quality is also very good as the audio is recorded with professional sound equipment in a very sterile environment with no background.

With lengths ranging from 20 to 40 minutes, the original recordings were unsuitable for training speech models and so a forced-alignment pipeline was developed to segment the recordings and their transcriptions into shorter parts (10-15 seconds in length). This work resulted in the creation of 260 hours of speech training data. An early version of this dataset was successfully tested 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.

In addition to the data from Český rozhlas, we collected 1454 hours of data from the Czech Parliament. The same processing pipeline was applied in order to make the data suitable for ASR training. A subset of 444 hours of this data was released publicly. A description of the dataset was published at LREC 2020 (Kratochvil et al., 2020). The paper can be found in Appendix A.

The combined dataset was used to train the ASR model that is in use as part of our subtitling demos.

2.2 Multi-Domain English Data

In order to build speech recognition systems which are robust to environmental noise and varying conversational styles, KIT has collected several speech training datasets from different domains such as telephone conversations, talks and lectures, read speech and broadcast news. We then merged all these into a single multi-domain dataset containing 1,622 hours of speech data. Table 1 shows the make-up of the dataset.

Dataset	**Domain**	**Hours**
Switchboard	Telephone Conversation	318
TED-LIUM	Lecture Presentation	453
Libri	Read Speech	363
Hub4	Broadcast News	148
ICSI Meeting	Meeting Discussion	76
Quaero	Broadcast News	188
WSJ	Dictation on Wall Street Journal	81
MIT Lectures	University Lectures	200
Total		1622

Table 1: Multi-domain speech recognition dataset.

In contrast to prior work, our composed multi-domain set is fairly balanced in the sense that it has no disproportionately large sub-set. Since the audio from Switchboard and Hub5'00 corpora was originally sampled at 8kHz, we converted it to 16kHz to have an uniform format.

This dataset was used as the basis for experiments contrasting the cross-domain performance of end-to-end versus traditional hybrid ASR systems. This work was published at the Life Long Learning for Spoken Language Systems Workshop colocated with ASRU 2019 (Nguyen et al., 2020).

3 MT Training Data

For most language pairs targeted by the project, some form of parallel data was already available, though not in the domain of auditing. In order to provide training data for the research, and integration systems required by the project, we have focused our efforts on: i) achieving high coverage by curating large scale multi-domain multilingual datasets that cover a majority of the 287 language pairs; ii) collecting new in-domain data corpora through targeted web crawling and sentence alignment.

3.1 OPUS Data

The OPUS collection[1] is a large collection of parallel corpora covering hundreds of languages pairs. The corpora cover a wide range of domains and genres, from translations of the bible to movie subtitles to parliamentary proceedings. UEDIN has sampled data from OPUS to produce several large multilingual corpora, each with a different emphasis.

3.1.1 The ELITR OPUS Corpus v1.0

For v1.0 of the ELITR OPUS Corpus (described in more detail in D1.1), we focused on covering a high proportion of the project's language pairs using data that was prioritized according to its domain (while there is no in-domain data in OPUS, the domains of some corpora are closer than others). The objective was to create a dataset from which we could build strong multilingual systems as early as possible in the project.

Drawing from the Europarl, EUBooks, and OpenSubtitles corpora (in that order), we sampled up to 1M sentence pairs per language pair, using upsampling for pairs with less data. This resulted in a corpus of 226M sentence pairs that was used to train the project's initial MT systems. While this corpus gave good language coverage, constraining the number of sentence pairs per language pair artificially limited translation quality for well-resourced languages. This corpus has now been superseded by v2.0, described shortly.

3.1.2 OPUS-100

In order to study the problem of massively multilingual translation, UEDIN produced the OPUS-100 corpus. Compared to the v1.0 corpus, OPUS-100 draws on a larger and more diverse set of languages and covers a greater range of writing scripts.

Following Aharoni et al. (2019), OPUS-100 is English-centric, meaning that all training pairs include English on either the source or target side. Translation for any language pair that does not include English is zero-shot or must be pivoted through English. The dataset is at a similar scale to Aharoni et al. (2019)'s, with 100 languages (including English) on both sides and up to 1M training pairs for each language pair. We selected the languages based on the volume of parallel data available in OPUS. When used to train many-to-many systems, the data covers 9,900 translation directions (with the majority being zero-shot).

Table 2 lists the languages, other than English, and numbers of sentence pairs.

Unlike the v1.0 corpus, we did not curate the data or attempt to balance the representation of different domains, instead opting for the simplest approach of downloading all corpora for each language pair and concatenating them. We randomly sampled up to 1M sentence pairs per language pair for training, as well as 2000 for validation and 2000 for testing. To ensure that there was no overlap (at the monolingual sentence level) between the training and validation/test data, we applied a filter during sampling to exclude sentences that had already been sampled. This was done cross-lingually, so an English sentence in the Portuguese-English portion of the training data could not occur in the Hindi-English test set, for instance.

OPUS-100 contains approximately 55M sentence pairs. Of the 99 language pairs, 44 have 1M sentence pairs of training data, 73 have at least 100k, and 95 have at least 10k.

[1] `http://opus.nlpl.eu`

Table 2: Numbers of training, validation, and test sentence pairs in the English-centric OPUS-100 dataset.

Language		Train	Valid	Test
af	Afrikaans	275512	2000	2000
am	Amharic	89027	2000	2000
an	Aragonese	6961	0	0
ar	Arabic	1000000	2000	2000
as	Assamese	138479	2000	2000
az	Azerbaijani	262089	2000	2000
be	Belarusian	67312	2000	2000
bg	Bulgarian	1000000	2000	2000
bn	Bengali	1000000	2000	2000
br	Breton	153447	2000	2000
bs	Bosnian	1000000	2000	2000
ca	Catalan	1000000	2000	2000
cs	Czech	1000000	2000	2000
cy	Welsh	289521	2000	2000
da	Danish	1000000	2000	2000
de	German	1000000	2000	2000
dz	Dzongkha	624	0	0
el	Greek	1000000	2000	2000
eo	Esperanto	337106	2000	2000
es	Spanish	1000000	2000	2000
et	Estonian	1000000	2000	2000
eu	Basque	1000000	2000	2000
fa	Persian	1000000	2000	2000
fi	Finnish	1000000	2000	2000
fr	French	1000000	2000	2000
fy	Western Frisian	54342	2000	2000
ga	Irish	289524	2000	2000
gd	Gaelic	16316	1605	1606
gl	Galician	515344	2000	2000
gu	Gujarati	318306	2000	2000
ha	Hausa	97983	2000	2000
he	Hebrew	1000000	2000	2000
hi	Hindi	534319	2000	2000
hr	Croatian	1000000	2000	2000
hu	Hungarian	1000000	2000	2000
hy	Armenian	7059	0	0
id	Indonesian	1000000	2000	2000
ig	Igbo	18415	1843	1843
is	Icelandic	1000000	2000	2000
it	Italian	1000000	2000	2000
ja	Japanese	1000000	2000	2000
ka	Georgian	377306	2000	2000
kk	Kazakh	79927	2000	2000
km	Central Khmer	111483	2000	2000
kn	Kannada	14537	917	918
ko	Korean	1000000	2000	2000
ku	Kurdish	144844	2000	2000
ky	Kyrgyz	27215	2000	2000
li	Limburgan	25535	2000	2000
lt	Lithuanian	1000000	2000	2000
lv	Latvian	1000000	2000	2000
mg	Malagasy	590771	2000	2000
mk	Macedonian	1000000	2000	2000
ml	Malayalam	822746	2000	2000
mn	Mongolian	4294	0	0
mr	Marathi	27007	2000	2000
ms	Malay	1000000	2000	2000
mt	Maltese	1000000	2000	2000
my	Burmese	24594	2000	2000
nb	Norwegian Bokmål	142906	2000	2000
ne	Nepali	406381	2000	2000
nl	Dutch	1000000	2000	2000
nn	Norwegian Nynorsk	486055	2000	2000
no	Norwegian	1000000	2000	2000
oc	Occitan	35791	2000	2000
or	Oriya	14273	1317	1318
pa	Panjabi	107296	2000	2000
pl	Polish	1000000	2000	2000
ps	Pashto	79127	2000	2000
pt	Portuguese	1000000	2000	2000
ro	Romanian	1000000	2000	2000
ru	Russian	1000000	2000	2000
rw	Kinyarwanda	173823	2000	2000
se	Northern Sami	35907	2000	2000
sh	Serbo-Croatian	267211	2000	2000
si	Sinhala	979109	2000	2000
sk	Slovak	1000000	2000	2000
sl	Slovenian	1000000	2000	2000
sq	Albanian	1000000	2000	2000
sr	Serbian	1000000	2000	2000
sv	Swedish	1000000	2000	2000
ta	Tamil	227014	2000	2000
te	Telugu	64352	2000	2000
tg	Tajik	193882	2000	2000
th	Thai	1000000	2000	2000
tk	Turkmen	13110	1852	1852
tr	Turkish	1000000	2000	2000
tt	Tatar	100843	2000	2000
ug	Uighur	72170	2000	2000
uk	Ukrainian	1000000	2000	2000
ur	Urdu	753913	2000	2000
uz	Uzbek	173157	2000	2000
vi	Vietnamese	1000000	2000	2000
wa	Walloon	104496	2000	2000
xh	Xhosa	439671	2000	2000
yi	Yiddish	15010	2000	2000
yo	Yoruba	10375	0	0
zh	Chinese	1000000	2000	2000
zu	Zulu	38616	2000	2000

To evaluate zero-shot translation, we also sampled 2000 sentence pairs of test data for each of the 15 pairings of Arabic, Chinese, Dutch, French, German, and Russian. Filtering was used to exclude sentences already in OPUS-100.

OPUS-100 was used for research conducted as part of WP4 (the work was published at ACL 2020 and is described in Zhang et al. (2020) and D4.2). The dataset is now publicly available as part of the OPUS collection.[2]

3.1.3 The ELITR OPUS Corpus v2.0

Version 2.0 of the ELITR OPUS corpus was designed to improve the language coverage of v1.0 and increase the volume of data for well-resourced language pairs.

Between the creation of versions 1.0 and 2.0, the JW300 corpus[3] was added to OPUS. The addition of this corpus, which alone covers over 300 languages, filled in most of the gaps in our language pair coverage. As a result, all ELITR language pairs (7 source x 42 target) are covered, with the exception of {Czech, French, German, Italian, Spanish, Russian}-to-Montenegrin. Since English-to-Montenegrin *is* included in JW300, the other six source languages are translatable into Montenegrin via pivoting.

The volume of data available from OPUS varies dramatically by language pair and by domain. Unlike in v1.0, we did not give precedence to any domain when sampling data. In order to reduce the dominance of high resource pairs / domains, we applied exponentially smoothed weighting (as used in multilingual BERT[4]) to domain and language pair choice when sampling sentence pairs.

In total, we sampled 1B sentence pairs (with replacement) from OPUS, which we then filtered to remove sentences that were present in the v1.0 test sets. Compared to v1.0, the resulting corpus is much larger: 975M sentence pairs compared to 226M. Table 3) shows the make-up of the dataset.

Subsets of this corpus have been used to train some of the MT systems currently in production as part of the PV platform (WP6). For instance, the subset of 233M sentence pairs that include English was used to train a one-to-many English-to-41 system, and similarly for German-to-40 and Czech-to-40 systems.

3.2 In-Domain Data

The OPUS collection covers a wide variety of domains and thus genres and our derived corpora are intended to be 'general-domain.' Collecting additional data for the auditing domain enables the possibility of building specialized models that perform better on in-domain text.

3.2.1 SAI Data

As reported in D1.1, UEDIN has collected in-domain data from all websites of the supreme audit organisations in EUROSAI to create a monolingual in-domain corpus. We used the list of EUROSAI members[5] to identify sites for crawling, and also crawled EUROSAI itself using a version of the ParaCrawl[6] crawling pipeline bitextor[7] We crawled PDF files as well as HTML, but at the time of creating the corpus PDF extraction was not yet supported by Bitextor and so were not included. We extracted text for all languages supported by the Moses[8] sentence splitting tools, giving data for a total of 24 languages with between 626 and 218,907 (median 14,182) sentences of text per language. See D1.1 for further details.

[2] http://opus.nlpl.eu/opus-100.php
[3] http://opus.nlpl.eu/JW300.php
[4] https://github.com/google-research/bert/blob/master/multilingual.md
[5] https://www.eurosai.org/en/about-us/members/
[6] http://www.paracrawl.eu
[7] https://github.com/bitextor/bitextor
[8] http://www.statmt.org/moses

	Czech	English	French	German	Italian	Spanish	Russian
Bulgarian	5962k	7499k	5540k	4237k	5195k	6580k	3727k
Croatian	4973k	6475k	4781k	3579k	4406k	5573k	3385k
Czech	-	8262k	6157k	4901k	6009k	7185k	3986k
Danish	4384k	6094k	5223k	4935k	5104k	5403k	2244k
Dutch	6214k	10038k	8630k	6726k	6967k	7907k	3712k
English	8262k	-	20566k	11949k	11030k	18090k	11002k
Estonian	4026k	4669k	3837k	3816k	3631k	4142k	2001k
Finnish	5312k	7377k	5685k	5316k	5389k	6233k	3042k
French	6157k	20566k	-	7841k	8033k	12934k	8573k
German	4901k	11949k	7841k	-	6785k	7418k	2857k
Greek	6164k	8623k	6828k	5903k	6300k	7661k	3756k
Hungarian	6566k	8440k	6494k	4981k	5992k	7118k	3966k
Irish	281k	590k	325k	325k	331k	329k	121k
Italian	6009k	11030k	8033k	6785k	-	8095k	3779k
Latvian	2329k	2866k	2770k	2744k	2360k	2381k	550k
Lithuanian	2516k	3066k	2838k	2775k	2425k	2467k	746k
Maltese	1482k	2007k	1959k	1917k	1497k	1494k	258k
Polish	6809k	8909k	6851k	5569k	6272k	7554k	4027k
Portuguese	5933k	10154k	7597k	6417k	6640k	8042k	3415k
Romanian	6527k	8935k	6476k	4936k	5805k	7533k	4013k
Slovak	3405k	4311k	3578k	3247k	3144k	3475k	1567k
Slovene	4604k	5521k	4699k	4060k	4093k	4786k	2516k
Spanish	7185k	18090k	12934k	7418k	8095k	-	8623k
Swedish	4598k	6678k	5506k	5097k	4814k	5421k	2443k
Albanian	943k	1206k	917k	845k	879k	954k	823k
Arabic	4692k	9218k	8222k	3130k	4183k	8917k	7715k
Armenian	283k	286k	286k	286k	275k	277k	290k
Azerbaijani	194k	309k	193k	225k	189k	187k	257k
Belorussian	64k	77k	69k	72k	68k	65k	75k
Bosnian	2558k	3089k	2249k	1695k	2058k	2783k	1691k
Georgian	385k	394k	383k	378k	365k	383k	390k
Hebrew	4605k	5329k	4234k	3035k	4036k	4953k	3311k
Icelandic	783k	978k	910k	834k	701k	811k	600k
Kazakh	51k	69k	56k	61k	58k	59k	59k
Luxembourgish	13k	20k	13k	20k	18k	14k	13k
Macedonian	1169k	1371k	1118k	978k	1026k	1231k	950k
Montenegrin	0k	65k	0k	0k	0k	0k	0k
Norwegian	2166k	2675k	2249k	2088k	2116k	2396k	1652k
Russian	3986k	11002k	8573k	2857k	3779k	8623k	-
Serbian	4977k	6643k	4902k	3225k	4207k	5996k	3328k
Turkish	5606k	7366k	5250k	3751k	4806k	6441k	3950k
Ukrainian	801k	983k	838k	795k	838k	873k	781k

Table 3: The ELITR OPUS v2.0 dataset.

3.2.2 ECA Data

UEDIN has applied a similar approach and extracted in-domain parallel corpora from the website of the European Court of Auditors (ECA)[9], which publishes reports in all 24 EU languages. We crawled reports from the ECA website in all EU languages in PDF format. After conversion to plain text (using macOS's Automator tool) we extracted sentence pairs for all language pairs where the source is one of Czech, English, French, German, Italian, and Spanish. The process involved a combination of translation (using a many-to-many NMT model trained on an EU-specific subset of the ELITR-OPUS v2.0 corpus) and Paracrawl's `bleualign` tool.

Table 4 shows the make-up of the dataset.

	Czech	English	French	German	Italian	Spanish
Bulgarian	93k	273k	106k	43k	14k	95k
Croatian	69k	181k	52k	25k	8k	48k
Czech	-	296k	90k	47k	16k	90k
Danish	34k	137k	51k	47k	30k	49k
Dutch	66k	304k	142k	105k	28k	130k
English	296k	-	454k	275k	68k	436k
Estonian	46k	138k	41k	33k	14k	42k
Finnish	97k	376k	125k	60k	17k	120k
French	90k	454k	-	85k	31k	253k
German	47k	275k	85k	-	19k	74k
Greek	89k	403k	178k	70k	22k	164k
Hungarian	99k	299k	95k	46k	14k	93k
Irish	-	-	-	-	-	-
Italian	16k	68k	31k	19k	-	34k
Latvian	22k	68k	21k	16k	15k	21k
Lithuanian	48k	155k	49k	34k	14k	49k
Maltese	112k	321k	129k	57k	21k	125k
Polish	119k	301k	101k	51k	15k	99k
Portuguese	94k	413k	198k	79k	33k	318k
Romanian	30k	94k	40k	23k	22k	41k
Slovak	224k	314k	101k	54k	18k	99k
Slovene	48k	103k	32k	23k	16k	31k
Spanish	90k	436k	253k	74k	34k	-
Swedish	108k	396k	142k	74k	20k	133k

Table 4: In-domain ECA dataset. Counts are 1000s of sentence pairs. Processing is still in progress for the language pairs that include Irish.

[9] `https://www.eca.europa.eu/en/Pages/ecadefault.aspx`

4 SLT Training Data

Compared to ASR and MT, datasets for SLT are scarce. One under-utilized source of data is the plenary sessions of the European Parliament. Building on an approach taken in recent work by Iranzo-Sánchez et al. (2020), we have processed and aligned large volumes of speech data with translations for all EU languages.

4.1 European Parliament Data

European Parliament (EP) is an extensive resource of multi-lingual parallel data. It not only contains texts for MT, but also voices of speakers and interpreters.

Iranzo-Sánchez et al. (2020) recently created Europarl-ST corpus. It contains speech-to-text translation data from plenary sessions between 2008 and 2011, because in this period the videos were published together with translations into all 23 EU languages. (There were only texts before 2008, and only transcripts in original languages, videos and simultaneous interpretations after 2011.) Europarl-ST contains only several language directions, excluding Czech. The authors included only the speeches for which they could automatically align audio and transcripts/translations, by matching video metadata with minutes, and by using automatic diarization and filtering by verbosity of transcripts. It may be used for multi-target speech-to-text translation, however, it does not include interpreters, and therefore it can not be used for for parallel multi-source speech-to-text translation.

There exist several corpora of simultaneous interpretations (SI) even from European Parliament (e.g. EPIC, EPIC-Ghent, EPTIC), but none of them are suitable for our research. These corpora are either not accessible for us, or are very size-limited, and contain only interpreters' transcripts without timing information and audios.

Therefore, we decided to create our own corpus of SI. We followed the approach of Europarl-ST for matching translations with video metadata, and for diarization. We downloaded the revised transcripts, translations and videos from all plenary meetings between 2008 and 2012, available on EP's website. The distribution of the languages spoken on plenary sessions is unequal. Table 5 contains the summary of language distribution based on metadata. The top languages are English (239 hours), German (110 hours) and French (100 hours). There are 15 hours of Czech. However, the metadata are inaccurate, they only estimate the reality.

We figured out in which dates the videos with speakers and SI and translations were available, see Table 6. We further focus on the period from 2008/9 to 2011/7.

At first, we decided to focus on English, German and Czech, so we downloaded the SI into these three languages. We processed them by automatic diarization, aligned the speeches with transcripts and translations, and processed them with ASR. We selected 170 hours of speeches, excluding the chair, who is not verbosely transcribed when chairing the session, where the automatic processing suceeded, for further analysis. See deliverable D4.2 for more details.

Furthermore, we selected 374 speeches, 10 hours, for held out dev-test set. We selected the same speakers for dev-test, as English-German part of Europarl-ST.

We processed dev-test manually. We corrected the automatic diarization, and transcribed the interpretations from Czech and German, and revised the transcript of English.

We plan to publish the corpus of speeches and SI. For every speech we include original audio with timestamped ASR transcripts, video, SI timestamped ASR, revised transcripts and translations, metadata and other information.

We contacted DG LINC Head of Strategy and Innovation Unit, and we received an authorisation to publish the transcripts of simultaneous interpreters, without their voices, which are protected as personal information. Therefore, we plan to publish links to EP's website, so the users of our corpus may download the SI on their own.

lang.	speeches	speakers	duration	English words
BG	283	25	7:15:45.32	70220
CS	587	38	15:54:57.16	167790
DA	353	22	8:39:6.87	99585
DE	3878	184	110:44:5.94	1123561
EL	941	56	24:13:28.22	218222
EN	7404	396	239:50:16.03	2272520
ES	1302	95	41:7:18.84	399147
ET	92	7	1:51:19.72	17566
FI	456	21	9:33:21.96	94404
FR	2674	193	100:54:37.06	1026783
GA	118	9	2:8:2.30	18554
HU	841	48	22:6:19.58	211309
IT	2081	117	53:38:31.42	495125
LT	288	18	5:43:17.77	54237
LV	124	12	2:54:6.71	27098
MT	82	6	2:38:13.95	22011
NL	1173	60	31:50:10.33	319709
PL	1787	93	41:53:59.32	403246
PT	930	38	25:51:51.14	235801
RO	1275	48	27:31:20.07	245186
SK	777	21	15:46:55.94	151648
SL	221	13	5:15:55.28	50426
SV	576	37	17:54:36.17	191228
XM	242	101	13:53:35.93	145367

Table 5: Language distribution of downloaded EP Plenary Session speeches according to metadata. XM language code is for unspecified or other language.

from	to	sessions	transcripts	translations	videos+SI
2008/01/14	2008/07/10	39	yes	yes	no
2008/09/01	2011/07/04	164	yes	yes	yes
2011/07/04	2012/12/13	78	yes	no	yes

Table 6: Availability of resources of EP Plenary Sessions at EP's website.

5 AM Training Data

The core portion of datasets for Automatic Minuting is the ELITR Corpus of Minuting, a set of meeting-and-minutes packages, created within the project. The other dataset is the qualitatively different ELITR-Europarl corpus, consisting of selected European Parliament Data in English paired with their corresponding minutes. We also use AMI (Mccowan et al., 2005) and ICSI (Janin et al., 2003) meeting corpora for testing purposes.

5.1 ELITR Corpus of Minuting

The ambitious task of creating automatic minuting of real project meetings demands data. Among available resources, there is a significant disproportion between coverage of different domains by the available open datasets which can be used for this purpose. For our experiments we primarily need project meetings, which are very rare in the available datasets: Although project meetings are being held all over the world thousands times a day, we can hardly use them, because the transcripts and minutes are mostly publicly unavailable.

These were the reasons for our decision to create our own corpus of meetings and minutes, which can be used as test data (or even training data) for automatic minuting. The corpus creation started with the project beginning and is growing and being developed during two years of the project duration.

The procedure is the following: (i) we collect meetings from our ELITR project meetings, as well as from other projects, if the participants gave their consents with the processing and foreseen future publishing of their data, (ii) we create ASR transcripts and manually correct them, (iii) we collect the existing (original) minutes for the meetings and create more minutes manually, with the help of annotators, (iv) using the specially created software, we manually align parts of corrected meeting transcripts to corresponding lines in the minutes. As the result, we get a large (the the biggest publicly available on the market) corpus of meetings with manually corrected transcripts, manually corrected minutes and the manually corrected transcript-to-minutes alignment. Furthermore, as the result of transcript-to-minutes alignment procedure, we get a secondary product, which can be used for evaluation of automatic minutes as well as for speech summarisation training: a set of relatively short pieces of speech paired to their one-line summaries.

We collect the minuting data for English and for Czech. By the end of 2020, we have gathered 144 hours of meetings in English and 64 hours of meetings in Czech. The length of the meetings varies from 10 minutes to more than 2 hours, with the median of roughly one hour. To get a better understanding about variability of summarization of the same speech, we have equipped many recordings with more than one manually created minutes. Table 7 summarizes the statistics about the collected meetings and annotations. In the first column, the duration and number of collected meetings recordings (shortened as “mtgs”) is given. The second column informs us about how many of these recordings have been manually corrected and equipped with minutes. The third column gives the amount of created minutes in terms of duration of the minuted meetings and the number of minutes created. These numbers are bigger than the length of the recordings due to the fact that many of the meetings (ca. 30% for English and more than 80% for Czech) are minuted more than once. In the last column we give the numbers for manual transcript-to-minutes alignment. This kind of annotation has begun recently, thus we don’t have too many minutes aligned to the transcripts yet. However, the annotation process is relatively fast, so we expect to catch up in the next few months.

	recordings	**corrected and minuted**	**total minuted**	**total aligned**
English	144h (134 mtgs)	106h (105 mtgs)	170h (172 mtgs)	45h (52 mtgs)
Czech	64h (63 mtgs)	55h (54 mtgs)	110h (109 mtgs)	25h (30 mtgs)

Table 7: Current statistics for the ELITR Minuting Corpus (“mtgs” means meetings).

The ELITR Minuting Corpus is not publicly released yet, we are still working on its creation. We will collect further meetings, supply minutes and provide the transcript-to-minutes alignments. Furthermore, we are working on the automatic deidentification of the transcripts and minutes, which should be also checked up manually before the corpus is published. We also expect that some meetings will have to be excluded from the public release of the ELITR Minuting Corpus for ethical reasons, and we don't make manual annotations on them.

5.2 ELITR Europarl Minuting Corpus

The European Parliamentary Meetings data are publicly available. For the automatic minuting task we need monolingual transcripts of meetings and the minutes. We decided to use the preprocessed transcripts from the European Parliament Proceedings Parallel Corpus (Koehn, 2005) and extracted the corresponding minutes from the European Parliament Website.[10] Out of 959 whole-day meetings in the EuroParl Corpus we obtained 427 whole-day meetings. We then automatically split them based on thematic units, which gave us about 5 times more individual meetings.

The extracted dataset represented more than 4000 hours of meetings with minutes, but more than a half of them appeared to be unusable (minutes are very short, uninformative or contain only names and names of the departments). Thus, the pairs have been semi-automatically filtered and we got ∼2000 of transcript-minute pairs. The corpus will be used as training and test sets for the planned minuting shared task.

It is important to note that EuroParl meetings and their minutes are very different in characteristics, objectives and domain from the ones we have collected from our own project meetings.

[10] https://www.europarl.europa.eu

6 ELITR Testset

In this section we report on the progress of `elitr-testset`, our collection of documents for comprehensive evaluation of our system. The design and composition of `elitr-testset` was described in the deliverable D1.4 (Year 2 Test Sets). Since then, further documents were added into the collection and the testset started getting used in regular evaluations of our systems-see the deliverable D6.3 on integration.

Some of the documents in `elitr-testset` belong to the *confidential* part of the testset due to copyright or privacy restrictions. Wherever possible, we publish the documents at the `elitr-testset` public repository[11], so that anyone can evaluate their systems in a mutually comparable way.[12] The current size of `elitr-testset` is summarized in the rest of this section, distinguishing documents usable for ASR evaluation, MT evaluation and SLT evaluation.

6.1 Indices

`elitr-testset` is an assorted collection of documents with topics ranging from administration to Mathematics. For reasons such as continuous evaluation of our ASR and MT workers to figure out the best system for our pipelines, we have found specific subsets of these documents to be more useful than the assorted collection of documents as a whole. To allow for an easy and automated evaluation in these different purposes, we introduce the concept of *indices*. Most of current indices are auto-generated based on the availability of the dataset for specific purposes and use cases. An *index file*,which can be generated by anyone, is technically just a list of documents to be considered for a specific evaluation task. The name of the index file then explains its purpose.

Currently we have, for example:

- an index that contains all files in the auditing domain that can be used to evaluate ASR,
- an index that groups all files that can be used when evaluating Czech to English MT,
- an index that groups all parallel EN-DE-CS documents.

6.2 ASR Evaluation Data Summary

Documents for ASR evaluation include sound files with the speech and it's reference transcripts. To evaluate latency, the transcripts are time-stamped at the word level. The total audio size, number of words and statistic information of the documents for ASR evaluation is listed in Table 8.

6.3 MT Evaluation Data Summary

`elitr-testset` contains multiple types and sources of parallel or multi-parallel documents. For simplicity of presentation, we disregard all these distinctions and report simply the total for each given pair of languages. We even disregard direction of the translation (i.e. our evaluation will suffer from the effect of translations). In the `elitr-testset` data itself we try to preserve all such metadata information at least in a non-formalized way so that the dataset can be later grouped as required. Some distinctions (such as our domains of interest) are captured formally by an index file which contains only the documents from the given domain as discussed in Section 6.1.

Some basic details about our auto-generated MT index files are present in Table 9. We aim towards inclusion of more MT language pair index files in the future, in-line with the growing size of `elitr-testset`.

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

[12] The issue of any potential overlap of training and test data remains open; the best strategy is to exclude `elitr-testset` sentences from the training data prior to training and to later exclude any training sentences from the test set prior to testing. (Both the training data and `elitr-testset` grow independently of each other and thus any overlaps can arise at any stage.)

Index	Lang	Duration	Words	Documents	Words per Document			
					Min	Max	Avg	Stddev
auto-czech-asr	cs	10:04:15	72408	5	3759	20592	14481.6	6658.59
auto-iwslt2020-devset	en	01:40:01	12602	7	171	4869	1800.29	1784.14
auto-iwslt2020-antrecorp	en	00:50:20	6634	37	79	292	179.30	46.65
auto-iwslt2020-khanacademy	en	00:18:04	4470	6	194	1584	745.0	554.45
auto-iwslt2020-consecutive	en	00:21:58	3207	2	1582	1625	1603.5	30.41
auto-iwslt2020-wgvat	en	01:17:14	8721	4	613	3451	2180.25	1173.73
auto-langtools-workshop	en	00:47:04	5185	5	480	1689	1037.0	507.37
auto-linguistic-mondays	en	01:16:29	12920	1	12920	12920	12920	NULL

Table 8: This table is extracted from auto-generated summaries of individual index files for `elitr-testset` commit-id **8ee98c622159afa38de994201eb0fc7c2c6d4297**. The prefix "auto-" in index names means that these indices were created automatically by considering all documents from a particular source.

6.4 SLT Evaluation Data Summary

Finally, some documents are useful for the evaluation of spoken language translation, i.e. they consist of the source audio, the reference transcript and the reference translation of the transcript into one or more languages.

The sizes of SLT evaluation data are summarized in the Table 10.

Index	Word Count	Source Lang	Target Lang
auto-mt-cs2en	27722202	cs	en
auto-mt-de2en	13241287	de	en
auto-mt-en2cs	32306014	en	cs
auto-mt-en2de	13439097	en	de
auto-mt-en2es	10769702	en	es
auto-mt-en2fi	4154627	en	fi
auto-mt-en2fr	6940244	en	fr
auto-mt-en2hr	10525282	en	hr
auto-mt-en2hu	3325413	en	hu
auto-mt-en2it	8231879	en	it
auto-mt-en2lt	515433	en	lt
auto-mt-en2lv	1187943	en	lv
auto-mt-en2mk	5883823	en	mk
auto-mt-en2nl	9031742	en	nl
auto-mt-en2no	4783017	en	no
auto-mt-en2pl	12038260	en	pl
auto-mt-en2pt	2345226	en	pt
auto-mt-en2ro	1756858	en	ro
auto-mt-en2ru	5843576	en	ru
auto-mt-en2sk	2029800	en	sk
auto-mt-en2sl	2528648	en	sl
auto-mt-en2sr	7393682	en	sr
auto-mt-en2sv	5518242	en	sv
auto-mt-en2uk	8162764	en	uk
auto-mt-es2en	10872636	es	en
auto-mt-fr2en	6940244	fr	en
auto-mt-it2en	8231879	it	en

Table 9: Automated summary of the **auto-mt-*** index files in `elitr-testset` at commit-id 8ee98c622159afa38de994201eb0fc7c2c6d4297. The word counts are rather big for regular test sets, due to the inclusion of all manually revised sentence-aligned pairs from Intercorp (`http://intercorp.korpus.cz/`) in the confidential part of `elitr-testset`. We plan to use only a small portion of it for regular testing.

Dataset	Source	Target	Duration	Word Count
auto-iwslt2020-devset	en	cs, de	01:40:01	12602
auto-iwslt2020-antrecorp	en	cs, de	00:50:20	6634
auto-iwslt2020-khan-academy	en	cs, de	00:18:04	4470
auto-iwslt2020-sao-consecutive	en	cs, de	00:21:58	3207
auto-iwslt2020-sao-wgvat	en	cs, de	01:17:15	8721

Table 10: `elitr-testset` data for the evaluation of the spoken language translation at commit-id 8ee98c622159afa38de994201eb0fc7c2c6d4297.

References

Roee Aharoni, Melvin Johnson, and Orhan Firat. Massively multilingual neural machine translation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3874–3884, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1388. URL https://www.aclweb.org/anthology/N19-1388.

J. Iranzo-Sánchez, J. A. Silvestre-Cerdà, J. Jorge, N. Roselló, A. Giménez, A. Sanchis, J. Civera, and A. Juan. Europarl-st: A multilingual corpus for speech translation of parliamentary debates. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8229–8233, 2020.

A. Janin, D. Baron, J. Edwards, D. Ellis, D. Gelbart, N. Morgan, B. Peskin, T. Pfau, E. Shriberg, A. Stolcke, and C. Wooters. The icsi meeting corpus. pages 364–367, 2003.

Philipp Koehn. Europarl: A Parallel Corpus for Statistical Machine Translation. In *Conference Proceedings: the tenth Machine Translation Summit*, pages 79–86, Phuket, Thailand, 2005. AAMT, AAMT. URL http://mt-archive.info/MTS-2005-Koehn.pdf.

Kratochvil, Polak, and Bojar. Large corpus of czech parliament plenary hearings. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 6365–6369, Marseille, France, May 2020. European Language Resources Association.

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.

Thai-Son Nguyen, Sebastian Stüker, and Alex Waibel. Toward cross-domain speech recognition with end-to-end models. *CoRR*, abs/2003.04194, 2020. URL https://arxiv.org/abs/2003.04194.

Biao Zhang, Philip Williams, Ivan Titov, and Rico Sennrich. Improving massively multilingual neural machine translation and zero-shot translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1628–1639, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.148. URL https://www.aclweb.org/anthology/2020.acl-main.148.

A Large Corpus of Czech Parliament Plenary Hearings

Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pages 6363–6367
Marseille, 11–16 May 2020

Large Corpus of Czech Parliament Plenary Hearings

Jonáš Kratochvíl, Peter Polák, Ondřej Bojar
Charles University
Faculty of Mathematics and Physics
Institute of Formal and Applied Linguistics
surname@ufal.mff.cuni.cz

Abstract

We present a large corpus of Czech parliament plenary sessions. The corpus consists of approximately 444 hours of speech data and corresponding text transcriptions. The whole corpus is segmented to short audio snippets making it suitable for both training and evaluation of automatic speech recognition (ASR) systems. The source language of the corpus is Czech, which makes it a valuable resource for future research as only a few public datasets are available for the Czech language.
We complement the data release with experiments of two baseline ASR systems trained on the presented data: the more traditional approach implemented in the Kaldi ASR toolkit which combines hidden Markov models and deep neural networks and a modern ASR architecture implemented in Jasper toolkit which uses deep neural networks in an end-to-end fashion.

Keywords: Czech, automatic speech recognition, dataset, speech corpus

1. Introduction

The field of automatic speech recognition (ASR) has been recently undergoing a methodological shift from hybrid model architectures towards end-to-end systems. These mostly neural-network-based architectures have recently been gaining momentum in their popularity and obtained state-of-the-art results on multiple popular speech corpora, e.g. Han et al. (2019) on test-clean set from LibriSpeech (Panayotov et al., 2015).
One particular downside of the end-to-end approaches is their requirement of an extensive amount of training data in order to produce competitive results in comparison with more traditional hybrid architectures.
At the same time, there has been active research in the field of fluent speech reconstruction (Fitzgerald et al., 2009; Cho et al., 2016; Klejch et al., 2017), which aims to produce a formal textual report of a particular recording. In reality, speakers often repeat certain words multiple times, correct themselves in the middle of a sentence or use informal language to describe their ideas. In many applications such as meeting reports, automatic subtitling of presentations or subsequent machine translation of the ASR output, we would like the transcripts to be a formal-language equivalent of what was said and not the exact copy of the actually uttered words.
In our paper, we present an extensive collection of Czech parliament plenary sessions which tries to contribute to both speech recognition per se as well as speech reconstruction. The corpus consists of 444 hours of spoken Czech together with the corresponding formal transcriptions as obtained by stenographers present at the meetings.
Our paper is organized as follows. In Section 2., we describe related literature and relevant work. In Section 3., we introduce our methodology used when constructing the corpus. Section 4. contains a detailed corpus description and its exploratory analysis. Section 5. presents the baseline results we obtained and we discuss them in Section 6. After a brief overview of our future plans (Section 7.), we conclude the paper in Section 8.

2. Related Work

There are relatively few publicly available speech-related resources for the Czech language.
The largest available Czech speech corpus is Prague DaTabase of Spoken Czech, PDTSC (Hajič et al., 2017)[1]. It comprises of approximately 122 hours of spontaneous dialog speech. The corpus is interesting not only for its size but also because it contains three layers of transcriptions: automatic speech recognition output aligned to audio (denoted as *z-layer*), literal manual transcript (*w-layer*) and finally speech reconstruction (*m-layer*).
There already exists a corpus of Czech parliament hearings published by the University of West Bohemia (Pražák and Šmídl, 2012), who collected the data from the year 2011 between February and August. The corpus comprises of 88 hours of transcribed speech. Another related Czech speech corpus is Vystadial 2016 (Plátek et al., 2016) which is a dataset comprising of 78 hours of telephone conversations in Czech.
With the same thematic as our corpus, there are some publicly available datasets for other languages. Multilingual dataset Europarl-ST (Iranzo-Sánchez et al., 2019) is a corpus of European Parliament debates in six languages. The most prominent parliament hearing corpus is a 2000-hours corpus of Finnish parliament (Mansikkaniemi et al., 2017). Another speech corpus is Althingi's Parliamentary Speeches of Icelandic parliament (Helgadóttir et al., 2017). There is also a relatively large corpus of 249 hours of Parliament hearings collected in Bulgarian parliament as described in Geneva et al. (2019).

3. Methodology

Our source data are available at the website of Chamber of Deputies[2] and they come in 15-minute audio segments. The corresponding texts for each of these audio files do not

[1] http://ufal.mff.cuni.cz/pdtsc1.0/
[2] http://www.psp.cz/eknih/2017ps/audio/2017/index.htm

match the exact spoken words as there is always some overlap between adjacent audio recordings. For this reason, we have first downloaded and concatenated all audio files for a particular hearing, removing the overlaps.
We also semi-automatically pre-processed the corresponding text transcriptions for individual parliament hearings. This pre-processing step includes mainly the removal of the text parts. We remove texts that do not correspond to the spoken audio. We also removed commentaries of some events during the hearing included by the stenographer. We normalized numerical values to text strings ("33" to "thirty three") and removed all non-speech related parts of the transcriptions.
The resulting concatenated audio files for each hearing are between 1 and 12 hours long. For segmentation of these long audio files to smaller parts, we used the Kaldi speech recognition toolkit (Povey et al., 2011). First, we trained a TDNN model (Peddinti et al., 2015) using our in-house Czech speech data and later used this model as the base model for the Kaldi audio segmentation script.[3] This method subdivides the text into shorter documents and performs decoding of the audio with a language model strongly biased towards the current document of interest. Note that these segments are related to the sound signal, not to sentence boundaries in the transcript.
After the segmentation step, we cleaned the data by discarding segments that did not match the decoded output. We have run this process in two iterations, always training a new model on the most recently segmented clean data and using it to generate new segments of the original files. The resulting audio segments are between 1 second and 44 seconds long.

4. Corpus Description

In this section, we provide an overview of the presented corpus. First, the way how the Czech parliament operates is described, then the obtained collection of the data, and finally, the corpus composition as we release it for the research community.

4.1. Czech Parliament

Czech parliament provides an audio recording of all their hearings together with the corresponding stenographic transcriptions starting from the year 2017. This period spans one election term of the Czech parliament.
During each parliament session, speakers usually take turns in communicating their opinions about a particular topic that is on the agenda. Speakers can also react to each other and ask for further clarification. In most cases, the hearing has a relatively predictable structure. First, the list of absent parliament members is read, and the proposed changes in the agenda are discussed and voted on. Second, the topic of the hearing is introduced, and members of the parliament take turns in describing their opinions and reacting to each other. Lastly, voting about the proposed changes takes place, and the hearing moves to the next topic on the agenda.

[3] https://github.com/kaldi-asr/kaldi/blob/master/egs/wsj/s5/steps/cleanup/segment_long_utterances_nnet3.sh

Audio hours	444
Audio files	191455
Number of hearings	85
Longest audio segment	44s
Shortest audio segment	0.5s
Unique speakers	212
Female speakers	48
Male speakers	164
Total words	3 029 646
Unique words	221 638
Time period	Nov. 2017 - Nov. 2019

Table 1: Corpus statistics

4.2. Corpus Statistics

In total, we have collected 85 parliament hearings over the considered period. These hearings are between 1 and 12 hours long.
The text transcriptions come from professional stenographers, who also revise the transcriptions and remove duplicate or repeating words, rewrite informal parts of speech to their formal equivalent and add non-speech marks, such as *laughter*, *background noise*, or a short description of the current situation in the parliament if it is noteworthy for the complete record of the hearing. During the plenary hearing, the stenographer on duty is seated directly next to the current speaker. There are between 8-10 stenographers present during each plenary hearing. Stenographers take turns every 10 minutes in transcribing the speech. Between their transcription turns, they polish the texts in order to adhere to the unified style specified by the Czech parliament.
We estimate that there is about 1-5% mismatch between the actual words the speaker said and the text that occurs in the transcript. We also note that the positive semantic correlation between the mismatch transcriptions and the actual words is very strong. The mismatch parts are, in most cases, more formally rewritten parts of the actual speech, which was too informally or colloquial.
We present the ASR corpus statistics in Table 1.

4.3. Final Corpus Composition

The final corpus consists of three versions of the data:

ASR Segments The audio is segmented to short parts, each complemented with the corresponding transcript in uppercase and without punctuation.[4] This version is directly usable for the training of the ASR acoustic model or end-to-end ASR.

ASR Transcript The plain text version of the full hearings, in the form suitable for ASR, i.e., uppercased, no punctuation, numbers spelled out, etc.

Plaintext Transcript This version of the transcript is sentence-oriented, true cased, and it also includes speaker identifiers, transcriber commentaries, and time information.

[4] If the full-length unsegmented recordings are required, please contact authors.

Section	Hours	Words
Training	433.5	2 966 148
Development	3.0	20 595
Test	7.5	42 903

Table 2: Sections of the released corpus.

The combination of ASR transcript and Plaintext transcript is useful for the training of sentence segmentation models. The direct output of an ASR system on the sound files together with plaintext transcript is useful for the training of speech reconstruction systems.
All three corpus versions were split into training, development, and test sets in the same way. The development set was taken to be a held-out part of the training set, whereas the test set was taken from a parliament hearing in a different election term.
The development set thus matches very closely the statistical distributions of the training data (but it, of course, contains unseen speech segments). The test set is deliberately made more challenging, avoiding speaker and topic overlap with the training data as much as possible.
We present the respective training, development and test set statistics in Table 2.

5. Experiments

In this section, we describe experiments with the proposed corpus of Czech Parliament Plenary Hearings. We conduct two experiments: first, we use the more traditional approach implemented in the Kaldi toolkit which uses the combination of Hidden Markov model (HMM), together with the Gaussian Mixture Model (GMM) and deep neural network. Second, we experiment with the end-to-end deep neural ASR architecture Jasper (Li et al., 2019).
We provide the results of our experiments in terms of word error rate (WER).

5.1. Kaldi

Kaldi (Povey et al., 2011) has been a popular speech recognition toolkit, especially for languages with limited data resources. The training of the model is done in two stages. First, we train the HMM-GMM model and make the alignments between speech and transcript of our training data. We then train a neural network on top of these alignments, which makes the acoustic model more robust. Kaldi is based on the underlying weighted finite-state transducers, which allows flexible incorporation of the n-gram language model that we use in our experiments as well.

5.1.1. Architecture Overview

For audio feature representation, we use 13 MFCC features (Davis and Mermelstein, 1980) together with their first and second-order derivatives, therefore our inputs for both the HMM-GMM and neural network training are 39-dimensional vectors. We first train the GMM-HMM model and make data alignments. As the neural network acoustic model, we use a combination of convolutional and time-delayed layer (Peddinti et al., 2015) neural network. For language modeling, we use the 4-gram language model, in the KenLM implementation (Heafield et al., 2013) trained on the corpus transcriptions.

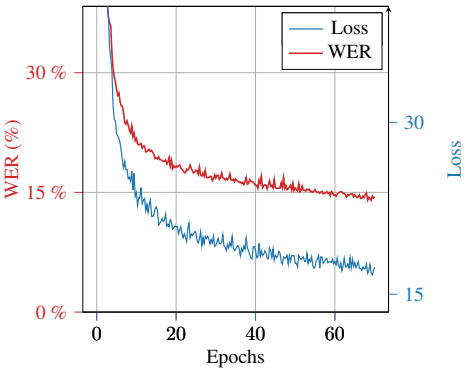

Figure 1: Jasper learning curve: the performance on the development set in terms of training loss and word error rate

5.2. CNN Jasper

Jasper is a family of end-to-end, deep convolutional neural network ASR architectures that unite acoustic and pronunciation models within one model. We decided to experiment with Jasper as it is a good example of contemporary end-to-end ASR solutions and because of its availability within the OpenSeq2Seq toolkit (Kuchaiev et al., 2018).

5.2.1. Architecture Overview

The input of the model are mel-filterbank features from 20 ms windows with 10 ms overlap. The output of the model is a probability distribution over characters from a custom vocabulary.
Jasper applies one pre-processing and three post-processing convolutional layers. Between these layers is the main part of the network, which consists of so-called "blocks".
The main part of the Jasper model consists of B blocks and R sub-blocks (the authors introduced convention where each Jasper model can be described as "Jasper BxR"). In our experiment, we use Jasper 10x5.
A sub-block comprises a 1D-convolution, batch normalization, ReLU activation, and finally, dropout. Each block input is connected to the last sub-block via residual connections. 1x1 convolution is used in order to project input channels to a different number of output channels. After each convolution layer, batch normalization is applied. The batch norm output is added to the output of the batch norm layer in the last sub-block. The sum is passed to the activation and dropout and produces the output of the current block.
Residual connections inspired by DenseNet (Huang et al., 2017) are employed to enable such a deep network to converge.

5.2.2. Training

As previously stated, we use the OpenSeq2Seq toolkit to build and train our end-to-end ASR model. Specifically, we

Method	Decoding	Dev WER	Test WER
Jasper	Greedy decoding	11.59 %	14.24 %
	Beam search	10.57 %	14.25 %
	Beam search LM	9.09 %	9.93 %
Kaldi	Beam search LM	6.64 %	7.10 %

Table 3: Experimental results of Kaldi a Jasper models on development and test sets.

make use of an existing implementation and configuration as provided by Jasper authors directly. Its main advantages include reduced training time and the implementation of mixed-precision training (all weights are stored in float16 while weights updates are computed in float32).
We altered the output vocabulary and extended it with characters used in the Czech language (characters with acute, caron, ring). Because of limited time and resources, we trained our model only for 70 epochs (instead of 400 as in the original paper). According to OpenSeq2Seq toolkit documentation (NVIDIA, 2018), authors indicate that training for 50 epochs should still yield acceptable results (for LibriSpeech dev-clean set it ought to be under 5% WER versus 3.61% after 400 epochs). We keep other network parameters unchanged.
We trained the network for three days and 20 hours on 8 Quadro P5000 GPUs. The progress and quality of the training were continuously checked on the development set, see Figure 1.

5.3. Results

We have used 433.5 hours of the training data to performs the experiments and evaluate the models on the development and test set.
The development set was randomly extracted from the training data, so the speaker's independence between the training and development set is not explicitly ensured. (It is possible but unlikely that a rare speaker was exclusively selected for the development set.)
The test set was especially chosen with the intention to reduce the possibility of speaker overlap with the training data, to arrive at a more realistic setting. With unseen speakers in the test set, any overfitting to the training data would be discovered.
We compare three methods of decoding: greedy decoding (for each time frame, the most probable character is selected), beam search and beam search with language model rescoring (Hannun et al., 2014).
In this experiment, we use the 4-gram language model in the KenLM (Heafield et al., 2013) implementation trained on the transcripts of the training set. Both, the plain beam search and the beam search witch language model rescoring use beam width of 256. The results can be seen in Table 3.

6. Discussion

From the results, we see that the Kaldi model slightly outperforms Jasper architecture. This may be mainly caused by the fact that end-to-end systems require even larger training data in order to match or improve over the hybrid architectures.
We have also used only n-gram language models in our experiments. In the future, we would like to add Transformer (Vaswani et al., 2017) as a language model, in a separate rescoring phase. This could substantially improve the results because the Transformer can capture longer contextual information than n-gram models.

7. Future Work

In the future, we want to process additional recordings and transcription from Czech parliament and senate using our existing data preparation pipeline.
We believe that we can extract an additional 1000 hours of audio data together with their transcriptions, which are already available. Moreover, as the parliament hearings take place each month, and there is a constant inflow of data at the rate of approximately 30 hours per month, so we would like to automate the whole dataset preparation pipeline and extend our dataset on the fly with each new hearing release.
Further, we would like to refine segmentation and alignment using the Jasper ASR model pre-trained on the current dataset. Al will allow us to segment recordings with respect to sentences rather than voice activity detection boundaries, enabling to train e.g., end-to-end ASR models with a built-in language model.
Another challenge of our interest is to balance the dataset with respect to the individual speakers (e.g., the president speaker opens every session) and also allow ensure gender balance, if desired.

8. Conclusion

We presented a new Czech speech corpus of 444 hours of Czech parliament plenary hearings. Further, we ran baseline experiments with speech recognition systems trained on the new corpus. We demonstrate that the proposed corpus is suitable for training both the traditional ASR models such as Kaldi, but also the state of the art end-to-end neural networks, yielding excellent results on benchmark development and test sets that we have prepared.
We release the corpus online for public use at:

`http://hdl.handle.net/11234/1-3126`

9. Acknowledgements

This research work was supported by the the project no. 19-26934X (NEUREM3) of the Czech Science Foundation and ELITR (H2020-ICT-2018-2-825460) of the EU research and innovation program.

10. Bibliographical References

Cho, E., Niehues, J., Ha, T.-L., and Waibel, A. (2016). Multilingual disfluency removal using nmt. In *Proceedings of the 13th International Workshop on Spoken Language Translation (IWSLT), Seattle, USA*.
Davis, S. and Mermelstein, P. (1980). Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE transactions on acoustics, speech, and signal processing*, 28(4):357–366.

Fitzgerald, E., Jelinek, F., and Frank, R. (2009). What lies beneath: Semantic and syntactic analysis of manually reconstructed spontaneous speech. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 746–754, Suntec, Singapore, August. Association for Computational Linguistics.

Geneva, D., Shopov, G., and Mihov, S. (2019). Building an asr corpus based on bulgarian parliament speeches. In Carlos Martín-Vide, et al., editors, *Statistical Language and Speech Processing*, pages 188–197, Cham. Springer International Publishing.

Hajič, J., Pajas, P., Ircing, P., Romportl, J., Peterek, N., Spousta, M., Mikulová, M., Grůber, M., and Legát, M. (2017). Prague DaTabase of spoken czech 1.0. LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

Han, K. J., Prieto, R., Wu, K., and Ma, T. (2019). State-of-the-art speech recognition using multi-stream self-attention with dilated 1d convolutions. *arXiv preprint arXiv:1910.00716*.

Hannun, A. Y., Maas, A. L., Jurafsky, D., and Ng, A. Y. (2014). First-pass large vocabulary continuous speech recognition using bi-directional recurrent dnns. *arXiv preprint arXiv:1408.2873*.

Heafield, K., Pouzyrevsky, I., Clark, J. H., and Koehn, P. (2013). Scalable modified Kneser-Ney language model estimation. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, pages 690–696, Sofia, Bulgaria, August.

Helgadóttir, I. R., Kjaran, R., Nikulásdóttir, A. B., and Guðnason, J. (2017). Building an asr corpus using althingi's parliamentary speeches. In *INTERSPEECH*, pages 2163–2167.

Huang, G., Liu, Z., Van Der Maaten, L., and Weinberger, K. Q. (2017). Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708.

Iranzo-Sánchez, J., Silvestre-Cerdà, J. A., Jorge, J., Roselló, N., Giménez, A., Sanchis, A., Civera, J., and Juan, A. (2019). Europarl-st: A multilingual corpus for speech translation of parliamentary debates. *arXiv preprint arXiv:1911.03167*.

Klejch, O., Bell, P., and Renals, S. (2017). Sequence-to-sequence models for punctuated transcription combining lexical and acoustic features. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5700–5704. IEEE.

Kuchaiev, O., Ginsburg, B., Gitman, I., Lavrukhin, V., Li, J., Nguyen, H., Case, C., and Micikevicius, P. (2018). Mixed-precision training for nlp and speech recognition with openseq2seq.

Li, J., Lavrukhin, V., Ginsburg, B., Leary, R., Kuchaiev, O., Cohen, J. M., Nguyen, H., and Gadde, R. T. (2019). Jasper: An end-to-end convolutional neural acoustic model. *arXiv preprint arXiv:1904.03288*.

Mansikkaniemi, A., Smit, P., Kurimo, M., et al. (2017). Automatic construction of the finnish parliament speech corpus. In *INTERSPEECH*, pages 3762–3766.

NVIDIA. (2018). Jasper — openseq2seq 0.2 documentation.

Panayotov, V., Chen, G., Povey, D., and Khudanpur, S. (2015). Librispeech: an asr corpus based on public domain audio books. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5206–5210. IEEE.

Peddinti, V., Povey, D., and Khudanpur, S. (2015). A time delay neural network architecture for efficient modeling of long temporal contexts. In *Sixteenth Annual Conference of the International Speech Communication Association*.

Plátek, O., Dušek, O., and Jurčíček, F. (2016). Vystadial 2016 – czech data. LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., Hannemann, M., Motlicek, P., Qian, Y., Schwarz, P., Silovsky, J., Stemmer, G., and Vesely, K. (2011). The kaldi speech recognition toolkit. In *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*. IEEE Signal Processing Society, December. IEEE Catalog No.: CFP11SRW-USB.

Pražák, A. and Šmídl, L. (2012). Czech parliament meetings. LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. (2017). Attention is all you need. In *NIPS*.