X Modal
X Cultural
X Lingual
X Domain
X Site
Global OER Network

Grant Agreement Number:	761758
Project Acronym:	X5GON
Project title:	Cross Modal, Cross Cultural, Cross Lingual, Cross Domain, and Cross Site Global OER Network
Project Date:	2017-09-01 to 2020-08-31
Project Duration:	36 months
Deliverable Title:	D3.4 – Early support for cross-lingual OER
Lead beneficiary:	UPV
Type:	Report
Dissemination level:	Public
Due Date (in months):	24 (August 2019)
Date:	
Status (Draft/Final):	Draft
Contact persons:	Javier Iranzo, Álex Pérez, Jorge Civera, Albert Sanchis and Alfons Juan

Revision

Date	Lead author(s)	Commments
5-Aug-2019	J. Iranzo, A. Pérez, J. Civera, A. Sanchis and A. Juan	first draft
25-Aug-2019	Colin de la Higuera	Added some comments / suggestions

Contents

List of Figures

List of Tables

Abstract

The main objective of WP3 is the construction of the analytics engine that provides the relevant knowledge required to drive the operation of the OER and social network. This includes cross-lingual issues in Task 3.3 (M12–M30), whose main goal is to extend the analytics engine with capabilities to deal with multi-lingual collections of OER. D3.4 is to report the work done in Task 3.3 from M12 (August 2018) to M24 (August 2019). During this period, the main goal of Task 3.3 is to provide early support for cross-lingual OER.

1 Introduction

The main objective of WP3 is the construction of the analytics engine that provides the relevant knowledge required to drive the operation of the OER and social network. This includes analysis of learning and testing, cross-lingual aspects, links with educational theories, affective computing, etc. In addition, there are two important aspects that are studied over the duration of the project:

1. Fine-grained indexation of educational videos by transcription tools; and
2. Investigation of multicultural, pedagogical and juridical issues, with particular care on privacy.

The first of these two aspects, and cross-lingual issues in general, are covered in Task 3.3 from M12 (August 2018) to M30 (February 2020). This deliverable, D3.4 – Early support for cross-lingual OER, is to report the work done in Task 3.3 from M12 to M24 (August 2019), i.e. mainly in Year 2 (Y2).

In connection to the purpose of D3.4, it must be noted that, as discussed in [1, Appendix C], the work on early support for cross-lingual OER already started during the last months of Year 1 (Y1). At that time we were developing the basic infrastructure and tools to support cross-lingual OER, X5gon-TTP. And with X5gon-TTP, we brought into production ASR (Automatic Speech Recognition) systems for automatic transcription of OER in the dominant languages of the official pilots, namely English, Spanish, Slovene and German. We used ASR systems from UPV background for English, Spanish and Slovene, though we also invested some effort to improve the Slovene ASR system, as well as to build a new German ASR system using state-of-the-art technology and having virtUOS, the official pilot from UOS, in mind. The technical details on these ASR systems were first reported in [1, Appendix C]. It is also important to note that all this previous work was limited to ASR systems; MT (Machine Translation) systems for automatic translation were left for Y2 (from M12 to M24, more precisely).

The work done in Task 3.3 from M12 to M24 has focused both on ASR, especially for English and Slovene, and MT, for language pairs covering not only the needs of the official pilots, but also an eventual expansion of the X5gon network to sites contributing OER in different languages. In this regard, we were aware that it was crucial for X5gon cross-lingual support services to end Y2 with full MT support for any pair of languages relevant to X5gon, maybe using English as a pivot language. For clarity and simplicity, a brief overview of the work done is provided in Section 2. The full story, with all relevant technical details, is provided in Section 3 for transcription, and in Section 4 for translation.

2 Overview of results

The overall progress achieved over the course of the project is summarized in Figure 1(a), for ASR, and Figure 1(b), for MT.

As can be observed in Figure 1(a), the main effort devoted in Y2 for ASR was to improve the English and Slovene ASR systems. We got consistent significant relative gains in WER for both languages. In the case of English, relative gains of 4% and 28% were achieved on the official VideoLectures.Net and

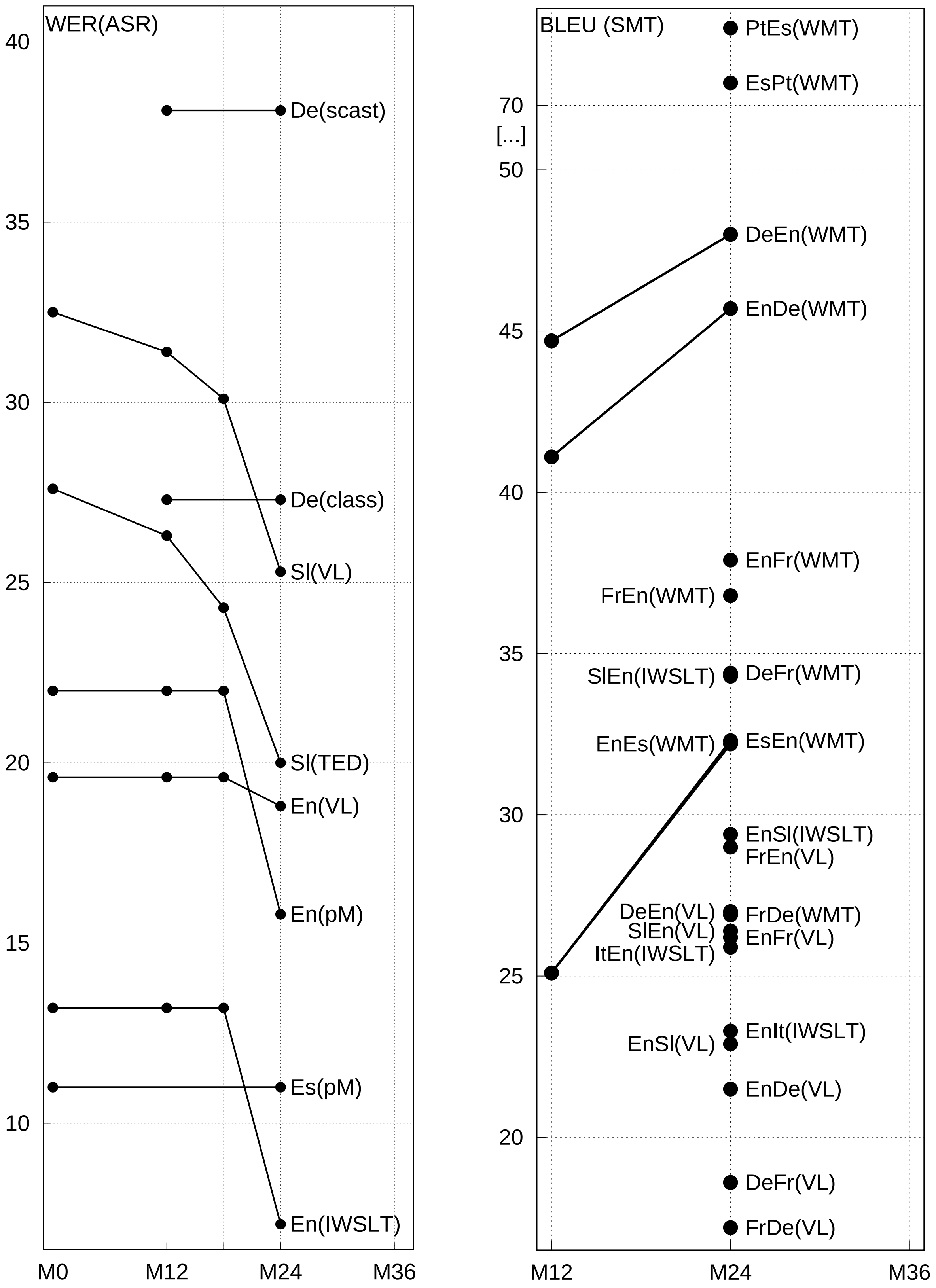

Figure 1: Progress for all languages in ASR on the left, given in terms of WER (the lower, the better) and in SMT on the right, in terms of BLEU (the higher, the better).

poliMedia pilots, respectively. This better performance was also confirmed with additional experiments on the official IWSLT 2013 test set over which a relative gain in WER of 45% was obtained. In absolute terms, it is observed that the performance of the English system is now below the threshold of 20% absolute WER points, which is often considered a clear indication of accurate transcriptions. In the case of Slovene, on the other hand, relative improvements of 19% and 24% were achieved on, respectively, the figures reported for VideoLectures.NET (VL) and SI-TEDx-UM (TED) in Y1. It is highly remarkable that we are now much closer to the 20% WER threshold for Slovene ASR; indeed it was crossed for TED.

As in the case of ASR, from Figure 1(b) we can easily spot the language pairs we dealt with in Y2 and the evaluation results we got. The Figure shows the evolution of the performance of the MT systems in terms of BLEU scores (the higher, the better) for language pairs involving English (En), German (De), French (Fr) Spanish (Es), Italian (It), Slovenian (Sl) and Portuguese (Pt) on the in-domain task, VideoLectures.NET (VL) and on well-known (out-domain) tasks that are widely used for comparison purposes by the MT research community (WMT and IWSLT). Generally speaking, the focus in Y2 has been on the deployment of a series of Neural MT (NMT) systems for language pairs of special interest in the project, and also for language pairs that will be certainly needed when extending the X5gon network to sites other than those of the official pilots. Moreover, as can be observed in Figure 1(b), most systems are the first systems deployed in X5gon for their corresponding language pairs. The only exceptions are those for German-English and English-German, with relative improvements of 7 and 11%, respectively, and that for Spanish-English, with a significant 30% relative increase. In brief, many of the systems deployed exhibit BLEU scores clearly above 35, or just below 35, which is a common reference for experts to consider them good enough for practical use. For systems showing scores below 30, which includes most in-domain evaluations on VL, more effort is still required. To end this overview of MT results, we refer the reader to Section 4.15, where comparative results with Google Translate are provided. In brief, X5gon MT systems are more or less on par with Google Translate for most language pairs, with the exception of Italian ↔ English, in which Google Translate is clearly ahead of X5gon, and Slovenian ↔ English and Portuguese ↔ Spanish, in which X5gon MT systems clearly outperform Google Translate. To us, being far ahead of Google Translate in key language pairs such as Slovenian ↔ English is a solid evidence that effective cross-lingual support for X5gon can only come from state-of-the-art MT systems *adapted* to the X5gon domain.

3 Transcription of OER

3.1 Transcription of German OER

3.1.1 Preliminary work

This section describes the ASR system developed for German during Y1 as was introduced in the Annex C of the deliverable D5.1 [1]. It has been specifically developed to transcribe German videos from the virtUOS pilot. The system is composed of two main separated models: the language model (based on words) and the acoustic model (based on phonemes). Additionally, a third (simpler) model is required to map words into phonemes. This last model, usually called *lexical model*, is used to join both acoustic and language models in order to obtain an automatic transcription.

In what follows we first describe the resources collected for training the different models, and then we introduce more technical details for each of them.

Compilation of Resources

Resources employed to train the German ASR system are described in this section. They are organised into those needed to train the acoustic model (transcribed speech), those needed to train the language model (plain text), and those for the lexical model (pronunciation dictionaries).

Resources for Acoustic Modelling

There are not many freely available speech corpora for training acoustic models. Indeed, for German we only found one corpus: *German Speech Corpus by Technische Universität Darmstadt* (GSC-TUDa) [2]. This corpus only contains 30 hours of annotated speech recorded using 5 different devices (about 158h in total), which is not enough to train a high quality acoustic model. Moreover, the recordings that can be found in GSC-TUDa are quite different from the recordings found in the virtUOS pilot. For theses reasons we decided to increase our resources by crawling publicly available data from the web. We collected all types of content: user generated videos, broadcast news, etc. The only requirement was to collect publicly available multimedia objects with subtitles.

Most of the subtitles found around the web can not be considered as phonetic transcriptions, which is required to train acoustic models. Moreover, in most cases the subtitles contain mistakes. In order to overcome these problems, the crawled data was forced-aligned using an initial acoustic model that was trained using the GSC-TUDa corpus. The resulted aligned data was then filtered using some heuristics based on the output of the alignment. As a result we managed to create a speech corpus containing about 716 hours of publicly available data crawled from the web. Table 1 sums up the basic statistics of all audio resources.

Table 1: Statistics of annotated speech resources for acoustic training of the German system.

Corpus	Duration(h)	Words(K)	Vocabulary(K)
Crawled Data	716	6150	180
GSC-TUDa	158	1161	4

Resources for Language Modelling

In contrast to what happens with acoustic resources, there are several monolingual and parallel text corpora which can be used to estimate a German language model (LM). In the case of parallel text corpora, only the German part is selected to estimate the model. Statistics of the LM resources are shown in Table 2. IWSLT De SMT refers to the monolingual part of the *International Workshop on Spoken Language Translation* (IWSLT) English-German statistical machine translation (SMT) task, while IWSLT De ASR 2013 Dev refers to a subset of the IWSLT 2013 German evaluation task used for development purposes.

Resources for Lexical Modelling

Instead of using a rule based approach, we decided to build the German lexical model using an automatic approach based on statistical models. For that purpose an initial supervised German pronunciation dictionary is needed. In our case, we used *The CELEX Lexical Database* (WEBCELEX) for German, which contains more than 310K words with their corresponding pronunciations [9].

Table 2: Statistics of German text resources for language modelling.

Corpus	Sentences(M)	Words(M)	Vocabulary(K)
Wikipedia [3]	65.2	642.1	8036.1
Europarl [4]	2.2	45.9	354.7
Common Crawl [5]	2.4	44.7	1313.6
News-Crawl [6]	2.0	29.6	661.6
Reuters [7]	0.5	17.6	280.5
Tatoeba [8]	0.3	2.6	86.5
IWSLT De SMT	0.2	3.2	120.9
IWLST De ASR 2013 Dev	0.001	0.02	3.6

System Description

In the following subsections, the models comprising the German ASR system (acoustic, language and lexical models) are described in detail.

Acoustic Model

The German ASR system performs speaker adaptation based on fCMLLR features [10]. A direct consequence of this is that it is composed by two hybrid HMM/NNs acoustic models [11]: a standard model and a fCMLLR model. In this setup the standard model is used in a first recognition pass to obtain a initial transcription, which is then used to perform fCMLLR normalisation over input features. The final transcription is obtained recognising the normalised features with the fCMLLR model. This scheme is usually referred as fCMLLR speaker adaptation [10, 12]. For the standard model a conventional feed-forward DNN is used, however for the fCMLLR model a BLSTM was used as in the Spanish system.

Estimation of the HMMs and DNNs was carried out using the transLectures UPV toolkit (TLK) [13], while BLSTMs were training using TensorFlow [14].

The characteristics for the final German acoustic models are the following:

- Standard model (1-pass)
  - 18867 tiedphoneme 3-state HMM with a 64-mixture component Gaussian per state.
  - A seven hidden layer DNN with the following architecture: 528 (48×11) input cells, 18867 output cells and 2048 cells in each internal layer.
- fCMLLR model (2-pass)
  - Same HMM topology than the standard model.
  - A five layer BLSTM with the following architecture: 48 input cells, 18867 output cells and 600 cells in each internal layer for each direction (600×2).

Language Model

The LM for German corresponds to a linear interpolation of several 4-gram models. Specifically, the six corpora in the upper part of Table 2 were used to estimate six independent 4-gram language models which were then interpolated. Interpolation weights were optimised using the *IWSLT De ASR 2013*

Dev set shown in the bottom part of Table 2. When the German ASR system was built, the amount of supervised data from the virtUOS pilot was scarce, so it was decided to use all of them for evaluation purposes. The choice of a subset of the IWSLT task as development, was based on the fact that the IWSLT task is composed of videos from TED talks [15], which have some similarities with educational videos. Therefore, it made sense to select this data for development purposes.

Apart from the conventional n-gram language model, an additional Recurrent Neural Network Language Model (RNNLM) was also trained for lattice rescoring using the Mikolov's toolkit [16]. That is, during the second decoding step (fCMLLR step) the decoder generates lattices containing the most promising hypothesis. The language model probabilities of these lattices are then rescored (replaced) by the ones obtained by interpolating the original 4-gram with the RNNLM. Most probable hypothesis is then selected as the final transcription. The IWSLT De SMT data set was used for training the RNNLM (see Table 2 for more details).

Lexical Model

The lexical model is a combination of the German dictionary described in Section 3.1.1 and a automatic grapheme to phoneme model estimated using the same dictionary. More precisely, the dictionary is first consulted to obtained the pronunciation(s) of a word. If no pronunciation is found, then the grapheme to phoneme model is used to obtain the most reliable pronunciation. The grapheme to phoneme model was estimated using the Sequitur G2P software [17].

3.2 Transcription of English OER

3.2.1 Preliminary work

During the first year of the project, transcriptions for English OER (VideoLectures.Net and poliMedia) were carried out using the MLLP's English ASR system, which was initially developed within the transLectures project [18]. The main characteristics of this initial ASR system were fully described in the Annex C of the deliverable D5.1 [1]. As a summary, this preliminary English ASR system was composed by two hybrid HMM/DNNs acoustic models as in the case of the German ASR system. Both acoustic models were trained using the same 2500 hours of speech data. Apart from the fCMLLR speaker adaptation, an additional step (3-pass) was carried out in which the fCMLLR DNN is adapted by performing conservative training using the output of the 2-pass. Also as in the case of the German ASR system, a conventional 4-gram model was built by interpolating several task-dependent 4-gram models and a classed-based RNN language model was trained with a small subset of in-domain data. During decoding a pruned version of the 4-gram model was used to generate lattices, which are then rescored using an interpolation of both 4-gram and RNN language models.

3.2.2 Work done in Y2

During the second year of the project a completely new English ASR system has been built from scratch. Compared with the preliminary system, this new system includes many improvements, which have resulted in significant gains on the quality of transcriptions. In what follows, the main characteristics of this new system are shown.

Resources for acoustic and language modelling

During Y2 the amount of speech data used for training the acoustic model has been drastically increased. In the previous system 2500 hours were used, while the new system has been trained with 5600 hours. This increase is partly due to the inclusion of new public available speech corpus, like

LibriSpeech or CommonVoice [19, 20], and partly due to the increase of the data crawled from the web. As in the case of the German ASR system, the crawled data was automatically aligned and filtered using the procedure described in [21]. Table 3 sums up the basic statistics of all audio resources.

Table 3: Statistics of annotated speech resources for acoustic training of the English system.

Corpus	Duration(h)
Crawled Data	3313.4
LibriSpeech [19]	959.7
TED-LIUM v3.0 [22]	453.8
CommonVoice [20]	242.5
SWC [23]	153.5
VideoLectures.NET [24]	109.9
Voxforge [25]	109.2
AMI [26]	96.1
EPPS [27]	79.4
ELFA [28]	48.3
VCTK [29]	43.9
poliMedia En [30]	2.4

Regarding LM resources, the corpora used are pretty much the same that the ones used in previous system with some few relevant differences. On the one hand, two new corpora have been added to the training: LibriSpeech and News-Discussions [19, 31]. On the other hand, the Wikipedia corpus has been updated to its last version. Statistics of the LM resources are shown in Table 4.

Acoustic model

For this new English ASR system, DNNs were replaced by Bidirectional LSTMs for acoustic modelling (BLSTMs) [42]. Since experimental results have shown that there is no improvement by using conservative training or fCMLLR, extra recognition steps were removed. Therefore, this new system is a more simple ASR system composed by a single BLSTM acoustic model trained with non-adapted features. Another important difference is that the dimension for the input vectors was increased. The original 16 MFCCs with derivatives (48 dimension vectors) were replaced by 80 MFCCs input vectors without derivatives.

Regarding the training procedure, the transLectures-UPV toolkit (TLK) was used to train a DNN-HMM model which was then used to bootstrap the BLSTM model [13]. In particular, BLSTM training consisted in a cross-entropy training procedure with a limited back propagation through time window of 50 frames, in a similar way than described in [42]. BLSTM training was carried out using TensorFlow [14]. The main characteristics for the resulting BLSTM model are:

- Output layer size (HMM topology): 16132 tiedphoneme 3-state HMM.
- 8 hidden layer BLSTM with 512 output cells per direction (1024 per layer).
- Input layer size: 80 MFCCs.

Language model

The LMs used in the new English ASR system were trained from scratch. Old corpora were preprocessed again, using a new preprocessing tool which takes into account abbreviations, number

Table 4: Statistics of English text resources for language modelling.

Corpus	Sentences(M)	Words(M)
Google Books count v2 [32]	-	294000.0
News-Discussions [31]	248.4	3649.5
Wikipedia [3]	149.9	2265.9
News Crawl [6]	53.1	1119.9
LibriSpeech [19]	40.4	803.6
GIGA [33]	22.5	616.8
United Nations [34]	12.9	334.1
HAL [35]	4.6	92.6
Europarl [4]	2.2	54.4
DGT-TM [36]	2.5	45.6
News-Commentary-v8 [37, 38]	0.2	5.5
WIT-3 [39]	0.2	2.7
COSMAT [40]	0.1	1.3
EuroParl TV [41]	0.1	1.2
VideoLectures.NET [24]	0.01	0.2
poliMedia [30]	0.002	0.036

conversions, among other new features. This new procedure was also applied to new corpora. Apart from the preprocessing, a new vocabulary was built. The size of this new vocabulary is 200K words. Any parameter tuning or topology decision, like the vocabulary selection, was taken on the basis of a development set. In order to create more robust LMs, we extended the old development set, which consisted only of VideoLectures.NET and poliMedia, to other tasks like: IWSLT or TED-LIUM. The final LM consisted of an interpolation of two different language models: a 4-gram LM and a LSTM LM.

The 4-gram model was trained using the conventional approach. For each corpus a specific 4-gram model was trained, and then they were interpolated in order to obtain the final 4-gram model. Regarding the LSTM LM, all corpora (with the exception of Google Books count) were first merged, and then random sentences were selected until reaching the amount of 1 billion words. This random selection was used to train the LSTM LM using Noise Contrastive Estimation (NCE) by means of the CUED-RNNLM toolkit [43]. In contrast to the old RNN LM, for this model the full vocabulary was modelled in the output layer. The final topology for the LSTM LM consisted of: an embedding layer of dimension 256, 1 hidden LSTM layer of dimension 2048 and an output softmax layer of dimension 200K. Lastly, the LSTM LM and the 4-gram were interpolated on the basis of the development set. The final weight for the LSTM LM in the interpolation was 0.79. Perplexities for both language model and the interpolation for the several development tasks are shown in Table 5. It is worth noting that TED-LIUM development was included in the training data for the old language model.

Apart from the new LM, the new English system includes another important novelty. In contrast to the preliminary system, in which a pruned 4-gram model was used to generate lattices which were then rescored, during Y2 a new decoder has been developed which is capable to directly interpolate on-the-fly the LSTM LM and the 4-gram during the regular decoding. We refer to this lattice-free decoding approach as one-pass-decoding. In addition to the simplification of the decoding pipeline, this new decoder is capable of avoiding cascade errors, that is, errors introduced by the pruned model during the decoding step which can not been fixed during the lattice rescoring step. More details

Table 5: Perplexities on the English ASR development sets for the new English LMs.

Task	4-gram	LM LSTM	Interpolation
VideoLectures.NET	273	258	200
poliMedia En	213	205	174
TED-LIUM v3.0 (legacy dev)	169	121	109
LibriSpeech (dev-other)	230	156	136
CommonVoice (valid-dev)	109	80	76

about the new one-pass-decoder can be found in [44].

Segmenter

Before running any decoding, input videos are automatically segmented into speech and non-speech segments. Each speech segment is then fed into the decoder in order to get the final transcription. The segmenter used in the preliminary English ASR system was based on Gaussian HMMs [45]. During this Y2 we have developed a new improved segmenter. This new segmenter is based on performing a first fast decoding step over the whole input video using DNN-HMMS. Based on silence boundaries in the decoder's output, the input signal is split into segments. These segments are then classified into speech and non-speech segment using the preliminary segmenter. More details can be found in [21].

Evaluation

In addition to the VideoLectures.NET evaluation set (originally developed for the transLectures [18]), the evaluation has been extended with two additional tasks: poliMedia English, an internal UPV evaluation set used to develop English systems for the poliMedia service [30], and the public test set defined for the IWSLT challenge in 2015 [46]. The corresponding development sets for each task were used also to tuning (grammar scale factor, pruning parameters, etc.). Statistics for these development and evaluation sets are shown in Table 6.

Table 6: Statistics (hours, running words and number of videos) for evaluation and development sets used in the development of the English ASR system

Task	Duration(h)	Words(K)	Videos
VideoLectures.NET test	3.2	34	4
poliMedia En test	2.9	28	27
IWSLT En tst2015	2.5	21	12
VideoLectures.NET dev	2.9	28	4
poliMedia En dev	2.9	28	27
IWSLT En tst2013	4.7	42	28

Results on development/evaluation sets for both the initial (with and without LM adaptation) and the new English ASR systems are shown in Table 7. It is worth noting that automatic LM adaptation to the task [47], which is used in the initial system, it is not available for the new system. The main reason is that this LM adaptation technique was developed for counts based models (for example 4-grams), and the new LM is mainly based on the use of a LSTM LM. The automatic adaptation of

LSTM LMs is still an open research question. Nonetheless, the new ASR systems outperforms the old system in all evaluation tasks, achieving relative improvements in the range of 4.1% ∼ 28.2% when the LM adaption is used in the initial system, and 7.8% ∼ 45.5% when it is not used.

Table 7: Comparison on WER% of initial and new English ASR systems for the English development/evaluation sets: VideoLectures.NET (VL), poliMedia En (pM) and IWSLT 2015 (IWSLT)

	Dev			Test		
	VL	pM	IWSLT	VL	pM	IWSLT
Initial (M0)	27.2	26.3	16.6	20.4	24.7	13.2
+ LM adaptation	26.2	22.6	–	19.6	22.0	–
New (M24)	23.3	18.9	8.5	18.8	15.8	7.2

3.3 Transcription of Slovenian OER

The Slovenian ASR System has been developed to transcribe Slovenian videos from the VideoLectures.Net pilot.

3.3.1 Preliminary work

At the beginning of the project, the Slovenian ASR system developed by UPV within the transLectures project was used. The main characteristics of this initial ASR system were described in the Annex C of the deliverable D5.1 [1]. Additionally, throughout this first period (M0 to M12 in Fig. 1(a)), a relative improvement of 3.4% in WER was achieved over this initial ASR system by applying Recurrent Neural Network language models (RNNLMs).

3.3.2 Work done in Y2

Throughout the second period, the Slovenian ASR system has been significantly improved by gathering new speech and text resources. Moreover, some improvements have been also obtained by applying the novel one-pass-decoding approach as has been previously described for the case of the English ASR system [44]. In what follows, we first outline the details of the new developed Slovenian ASR system along with the resources gathered for acoustic and language modelling and, after, we describe the experimental evaluation.

System Description

Acoustic Model

The Slovenian ASR system is composed by two hybrid HMM/NNs acoustic models. Both standard and fCMLLR models are multilingual feed-forward DNNs trained using Spanish and Slovene speech data. In particular, 800 hours of Spanish speech data plus 165 hours of Slovenian speech data summarized in Table 8 were used to train the acoustic models using the transLectures UPV toolkit (TLK) [13].

The main characteristics for these acoustic models are the following:

- Standard model (1-pass)
  - 14497 tiedphoneme 3-state HMM with a 64-mixture component Gaussian per state.

Table 8: Slovenian speech resources for acoustic training.

Corpus	Duration(h)	Running Words(K)	Vocabulary(K)
Gos [48]	39.7	360	37
GosVL [49]	21.9	169	24
RTVSLO [50]	77.0	634	66
VideoLectures.Net [50]	26.6	229	27
Total	165.2	1392	101

- A six hidden layer DNN with the following architecture: 528 (48 × 11) input cells, 14497 output cells and 2048 cells in each internal layer.

- fCMLLR model (2-pass):
  - 14983 tiedphoneme 3-state HMM with a 64-mixture component Gaussian per state.
  - A six hidden layer DNN with the following architecture: 528 (48 × 11) input cells, 14983 output cells and 2048 cells in each internal layer.

Language Model

As in the case of the English ASR system, new LMs were trained from scratch using the text resources summarized in Table 9. Moreover, a new vocabulary of 500K words was selected to build these news LMs. The final LM was an interpolation of two different LMs: a 4-gram LM and a LSTM LM.

Table 9: Slovenian text resources for language modelling.

Corpus	Sentences(K)	Running Words(K)	Vocabulary(K)
Europarl-v7 [4]	623	12559	135
ccGigafida [51]	7448	110098	1212
Gos [48]	26	360	37
Newscrawl2011 [6]	1000	18599	433
RTVSLO [50]	129	14399	309
slwac2.0 [52]	50847	778982	5446
TED [53]	54	339	39
VideoLectures.Net [50]	10	229	27
Wikipedia [3]	1804	22330	776
Wit3 [54]	15	200	29
Total	61956	958087	5962

The 4-gram model was built as a linear interpolation of ten 4-gram LMs trained for each one of the corpus summarized in Table 9. The interpolation weights were tuned to optimize the perplexity of the VideoLectures.Net development set (statistics of this corpus are summarized in Table 12). The optimal weights for each 4-gram model are summarized in Table 10.

In the case of the LSTM LM, all the text resources were used to train it using Noise Contrastive Estimation (NCE) by means of the CUED-RNNLM toolkit [43]. In contrast to the LSTM LM used in the first period, in this case the full vocabulary (500K words) was modelled in the output layer.

Table 10: Interpolation weights of the ten 4-gram LMs used to build the Slovenian 4-gram LM.

4-gram model	Weight [%]
slwac2.0	37.48
RTVSLO	37.17
VideoLectures.Net	18.95
Gos	3.59
ccGigafida	1.71
Wikipedia	0.40
Wit3	0.24
Europarl-v7	0.19
Newscrawl2011	0.18
TED	0.09

Thus, the final topology for the LSTM LM consisted of: an embedding layer and a hidden LSTM layer of dimension 1024 and an output softmax layer of dimension $500K$. Finally, the LSTM and the 4-gram LMs were interpolated to optimize the perplexity of the VideoLectures.Net development set. The final weight for the LSTM LM in the interpolation was 0.67. Perplexities for each LM over the VideoLectures.Net development and test sets are shown in Table 11. An Out-of-Vocabulary (OOV) rate of 0.7% and 1.7% for the development and test sets, respectively, was achieved with the new lexicon size of 500k words.

Table 11: Perplexities of the UPV's Slovenian LMs on the VideoLectures.Net dev and test sets.

Model	dev	test
4-gram	473	666
LSTM	319	401
Interpolation	284	363

Decoding

In the first period, the speech decoding process was performed following a three-pass decoding setup. The speaker-independent acoustic standard model was used primarily to obtain a transcription which in conjunction with a simple "target" Hidden Markov Model allowed for the transformation of acoustic features into speaker-adapted features [10, 12]. In the second pass, the speaker-adapted features were used along with the fCMLLR model to produce word-lattices. Both recognition steps were carried out using a pruned version of the 4-gram Slovenian LM to allow for very fast decoding. In a third final step, a LSTM LM was used for N-best rescoring.

In the second period, as was metioned before, a more efficient two-pass decoding setup has been applied [44]. Under this setup, the N-Best rescoring is avoided by using directly an interpolation of the 4-gram and the LSTM LMs during the second decoding pass.

Evaluation

In addition to the VideoLectures.NET evaluation set, the Slovenian ASR system was also evaluated on the publicly available SI-TEDx-UM corpus based on TEDx Talks [53]. The main statistics for these development and evaluation sets are shown in Table 12.

Table 12: Statistics for the Slovenian development (dev) and test sets.

Set	Corpus	Dur.(h)	#Spk.	Running Words(K)	Voc.(K)
Dev	VideoLectures.Net [50]	2.9	4	27.4	5.2
Test	VideoLectures.Net [50]	3.2	4	23.2	5.9
	SI-TEDx-UM [53]	3.2	13	26.9	7.0

The WER achieved on the evaluation sets for the Slovenian ASR systems developed during the first and second period are shown in Table 13. As can be observed, the new Slovenian ASR system outperforms the initial ASR system in both evaluation tasks achieving significant relative improvements of 19.4% and 24.0% on the VideoLectures.NET and SI-TEDx-UM evaluation tasks, respectively.

Table 13: WER of the Slovenian ASR system throughout the project.

Period	System	VideoLectures.Net	SI-TEDx-UM
Y1	Initial ASR System	32.5	27.6
	+ LSTM-LM NBest Rescoring	31.4	26.3
Y2	+RTVSLO speech data	30.1	24.3
	+New (N-gram and LSTM) LMs	28.1	22.6
	+GosVL speech data & one-pass LM decoding	25.3	20.0

4 Translation of OER

This section describes the Neural Machine Translation (NMT) systems updated or developed in Y2 of X5gon for the following language pairs: {German, Spanish, French, Italian, Slovenian} ↔ English, German ↔ French and Portuguese ↔ Spanish.

Our system architecture is based on the state-of-the-art Transformer model [55]. In order to train our systems, we have used two configurations, Transformer Base and Transformer Big, both with 6 encoder/decoder blocks, a self-attention mechanism, and the following features:

- **Transformer Base**: Embedding dimension 512, hidden layer size 2048 and 8 attention heads.
- **Transformer Big**: Embedding dimension 1024, hidden layer size 4096 and 16 attention heads.

The Big configuration has been shown to achieve better results, but it requires more data to properly estimate its parameters, and is harder to train. We have also experimented with training systems with more than 1 GPU, filtering techniques and backtranslations.

For each of those language pairs, we first present the resources used to train the MT systems. Secondly, we describe the configurations of the NMT systems. Finally, we evaluate the developed MT systems in terms of BLEU score [56] on an out-domain task obtained from a shared task in WMT or

IWSLT and on the in-domain task Videolectures.NET (VL), whenever the test set is available in the following languages: English, German, French, Spanish and Slovenian [?]. It is important to note that the VL task uses a very specific technical language, so that it becomes a challenging translation task.

Furthermore, we have also included comparative results providing those of Google's Machine Translation, in order to compare our X5gon results to those of an already existing mainstream provider.

4.1 German-English

The data published for the News Translation Shared Task of the WMT 2018 competition [57] was selected in order to train the German-English system. The data includes 4 well-established corpora (news-commentary, europarl, commoncrawl and rapid2016) as well as the recently released paracrawl corpus. This last corpus was collected by crawling a set of possibly noisy parallel web pages. This introduces the need of using filtering and data selection techniques to discard noisy sentences that could harm the performance of the NMT system. Table 14 shows statistics of the above-mentioned corpora.

Table 14: Statistics of the WMT German-English dataset.

Corpus	Sentences(K)	Words(M)		Vocabulary(K)	
		De	En	De	En
News-Commentary v13 [58]	284.2	6.4	6.2	302.8	182.5
Europarl v7	1920.2	44.6	47.9	649.0	304.8
Common Crawl	2399.1	47.0	51.4	2733.7	1718.9
Rapid 2016 [59]	1329.0	22.1	23.0	674.8	387.7
Paracrawl [60] v1	36351.6	450.7	478.8	14054.0	10353.1
Paracrawl v3	31358.2	465.2	502.9	14067.6	9905.3

The data was filtered using a language model approach described in [61], and we selected the 10M sentences with best score. Additionally, we included an additional set of 20M synthetic sentences produced using the Backtranslation approach [62]. We compare the results of the Transformer Base model with a second Transformer model that follows the same configuration, but trained using 3 GPU, and therefore it uses a batch size that is 3 times bigger than the previous one. Additionally, this second system was trained with longer sentences (maximum sentence length of 100 words, compared with 75 words of the previous one). Lastly, we also report results for a Tranformer Big model trained with 8 GPU[1]. The data setup for this last system is slightly different from the other two. First, we extracted 10M sentences using Cross-Entropy Filtering [63] from the Paracrawl v3 corpus, and used all data from the other corpora. Secondly, we used an additional set of 24M backtranslated sentences, and we applied noise to the source side of all backtranslations [64].

In order to evaluate our systems, we have elected to use a set of standard sets from the news translation task of the WMT competition, using newstest2015, newstest2017 and newstest2018. Table 15 shows the results obtained by the German-English MT systems.

The system trained with 3 GPU obtains improvements of around 1.0 BLEU compared with that of 1 GPU system. This improvement can be attributed almost entirely to the increase in batch size. The only other difference between the two systems is the maximum sentence length, and there are almost no sentences in the test set whose length is longer than 75 words, so we believe that the effect

[1] Because only 4 physical 4 GPU were available, we simulated the equivalent batch size of using 8 GPU with the gradient accumulation technique. As the only difference between the two approaches is a slower training when using gradient accumulation, we report the GPU-equivalent batch size used, disregarding the actual number of physical GPUs available

Table 15: Evaluation results of the German-English MT systems on the WMT task.

System	newstest 2015	newstest 2017	newstest 2018
Transformer Base	34.3	35.9	44.7
Transformer Base, 3 GPU	35.3	36.9	45.9
Transformer Big + Noise, 8 GPU	37.2	39.4	48.0

of this setting will be minor. The Transformer Big obtains further significant improvements over the previous best system, with an increase of 1.9, 2.5 and 2.1 BLEU, respectively.

This latter system was also assessed on the VL task obtaining 27.0 BLEU, that is far from that obtained in the out-domain task. This is due not only to the language specificity of VL and the lack of a large in-domain training set, but also to the stopping training criteria based on the out-domain task.

4.2 English-German

The resources chosen to build the English-German system are the same as those of the German-English MT system, as the language pair involved is the same. The data has been described in Section 4.1, and we have followed the same setup in this language pair, by selecting 10M sentences using filtering. details are shown in Table 14.

We present the results of the two systems developed for English-German, a Transformer Base and a Transformer Base model trained with 3 GPU. The second system is trained with a maximum sentence length of 100. Additionally, the second model was trained by augmenting the 10M parallel sentences with 18M backtranslations. The backtranslations were generated with the Transformer Base German-English model of Section 4.1. We have also used those backtranslations in order to train a Transformer Big model with 8 GPU.

Table 16 shows the results obtained by the English-German MT systems on the news translation WMT dataset. We observe how the 3 GPU model is able to obtain improvements of 2.0 BLEU in newstest2015 and newstest2017, following a similar pattern as the German-English system of Section 4.1. This increase is even bigger in newstest 2018, with a 4.1 BLEU improvement. The combination of a bigger batch size with the additional 18M backtranslations makes it hard to isolate the individual contribution of each change to the overall improvement, but based on previous experience, it is very likely that both changes have significantly contributed to the improvement. The Transformer Big model obtains additional improvements of 0.7, 0.7 and 0.5 BLEU, respectively. This increase is smaller than that observed in the German-English case, perhaps due to the smaller amount of backtranslations used.

As a future work, in order to reduce this difference, we could train the Transformer Big model using additional backtranslations as well as adding noise.

Table 16: Evaluation results of the English-German MT systems on the WMT task.

System	newstest 2015	newstest 2017	newstest 2018
Transformer Base	29.1	27.4	41.1
Transformer Base, 3 GPU + Backtrans.	31.1	29.4	45.2
Transformer Big, 8 GPU + Backtrans.	31.8	30.3	45.7

Similarly to the English-German pair, the BLEU figure on the VL task is 21.5, that again due to the domain mismatch is far from those figures achieved in WMT.

4.3 Spanish-English

We will now describe the data used for the Spanish-English language pair. We have 3 distinct type of corpora. The first consists in 6M pseudo in-domain data for OER. This is the data that was used to train our previous phrase-based SMT systems. We also have a series of general domain corpora (Common Crawl, EUbookshop, EU-TT2, EUTV and UN) [65] as well as a small in-domain corpora from the poliMedia repository. Table 17 shows statistics of each corpus.

Table 17: Statistics of the data sets used to train the Spanish-English MT systems.

Corpus	Sentences(K)	Words(M)		Vocabulary(K)	
		Es	En	Es	En
pseudo in-domain data	6005.7	144.1	133.4	820.7	756.2
Common Crawl	1 845.3	43.5	40.8	1555.2	1371.5
EUbookshop	5215.5	136.8	121.0	2203.1	2052.7
EU-TT2	1039.9	23.0	21.2	223.7	202.6
EUTV	180.5	1.8	1.9	70.5	56.9
UN	11196.9	366.1	320.1	668.2	651.7
poliMedia	150.0	2.3	2.4	122.5	88.2

We present the results for two Spanish-English systems, a Transformer Base model, and a Transformer Big model trained with 3GPU. The first system was trained using the 6M pseudo in-domain data. The second system has been trained using all the available data from the general-domain and in-domain corpora, and using a 3 GPU machine.

The Spanish-English systems have also been evaluated using a set of standard test sets from the news translation shared task of the WMT competition, as test sets are also available for this language pair. In this case, we use newstest2012 as development set and newstest2013 as test set. Table 18 shows the results of the Spanish-English models.

Table 18: Evaluation results of the Spanish-English MT systems on the WMT task.

System	dev BLEU	test BLEU
Transformer Base	27.2	25.1
Transformer Big, 3GPU full dataset	34.7	32.3

The Transformer Base model obtains 27.2 BLEU in the dev set and 25.1 BLEU in the test set. The Transformer Big model obtains 34.7 BLEU in the dev set and 32.3 BLEU in the test set, which represents an improvement of 7.5 BLEU and 7.2 BLEU, respectively. When comparing this with the results of Section 4.4, it is likely that the big improvement in BLEU is thanks mostly to the additional data, and not to the change from Base to Big model. As a future work, more bilingual data can be collected, as well as producing backtranslations.

In this language pair, the BLEU score on the VL task is 36.4 that is higher than that achieved in the out-domain WMT task. This is clearly explained by the abundance of pseudo in-domain data for OER in this language pair.

4.4 English-Spanish

The resources chosen to build the English-Spanish system are the same as those of the Spanish-English MT system, as the language pair involved is the same. The details are shown in Table 17.

We present the results for two Transformer Base models. As in the previous case, the first Transformer Base model is trained using a single GPU. The system was trained using the 6M pseudo in-domain data. The second system has been trained using all the available data from the general-domain and in-domain corpora, and using a 3 GPU system. Table 19 shows the results obtained by the English-Spanish MT systems.

Table 19: Evaluation results of the English-Spanish MT systems on the WMT task.

System	dev BLEU	test BLEU
Transformer Base	26.6	25.1
Transformer Base, 3 GPU full dataset	35.0	32.2

Following the trend of the Spanish-English systems of Section 4.3, the first Transformer Base obtains 26.6 BLEU in the dev set and 25.1 BLEU in the test set, whereas the 3 GPU model obtains 35.0 and 32.2 BLEU, respectively. This represents an improvement of 8.4 and 7.0 BLEU, respectively. Apart from the issues mentioned in the Spanish-English case, as future work, we will assess a Transformer Big architecture for this language pair, so further improvements are straightforward.

As in the Spanish-English pair, the BLEU score on the VL task is significantly higher than that achieved in the out-domain WMT task, 39.4 BLEU points.

4.5 French-English

The data used for training French-English systems is the WMT14 News Translation Shared Task French-English data. This is a well-known dataset that is frequently used in order to compare results in the literature, so it allows us to measure our progress compared with other research teams. This dataset contains two medium sized corpora (europarl and commoncrawl) as well as two significantly larger corpora, undoc, which is a collection of UN documents, and the Gigaword corpus, a collection of news text data with more than 20M parallel sentence pairs. Table 20 shows the statistics of the WMT14 dataset.

Table 20: Statistics of the data sets used to train the French-English MT systems.

Corpus	Sentences(K)	Words(M)		Vocabulary(K)	
		Fr	En	Fr	En
Common Crawl	3244.2	76.7	70.7	2081.8	1918.2
Europarl-v7	2007.7	52.5	50.3	417.8	311.9
GIGA	22520.4	672.2	575.8	6899.5	7029.5
News Commentary	183.8	4.7	4.0	175.9	146.4
UN	12886.8	354.2	316.5	2548.7	2079.8

We have trained 2 models for this language pair, a Transformer Base and a Transformer Big model using 3 GPUs. All models had a maximum sequence length of 100 tokens.

Following the setup of the WMT14 competition, we have used newstest2013 as the dev set, and newstest2014 as the test set. The results obtained by the French-English MT models are shown in Table 21.

Table 21: Evaluation results of the French-English MT systems on the WMT task.

System	dev BLEU	test BLEU
Transformer Base	33.1	36.8
Transformer Base, 3 GPU	33.0	36.8

In this case, both systems show similar performance, with 33.1 and 36.8 BLEU in the dev and test sets. This result is different from other language pairs, where an increase in batch size also meant an improvement in translation quality. Further experiments using the Transformer Big configuration as well as bigger batches could be a way of improving results. As already mentioned, a straightforward, although computationally expensive way to improve results, is to generate synthetic backtranslations for those language pairs that do not have them.

The BLEU score achieved on the VL task with 29.0 points is lower than that on the out-domain WMT task. As happened in the German pairs, the limited access to in-domain training data harms the performance of the system on the VL task.

4.6 English-French

The resources chosen to build the English-French system are the same as those of the French-English MT system, as the language pair involved is the same. The details are shown in Table 20.

We have trained a Transformer Base as well as a Transformer Big model, this one trained using 3 GPUs. The models were trained with a maximum sequence length of 75. In a similar way to the French-English case, we used newstest2013 as dev set, and newstest2014 as test set. Table 21 shows the results obtained by the English-French systems.

Table 22: Evaluation results of the English-French MT systems on the WMT task.

System	dev BLEU	test BLEU
Transformer Base	30.9	35.2
Transformer Big, 3 GPU	33.6	37.9

The Transformer Base obtains 30.9 BLEU in the dev set, and 35.2 BLEU in the test set, whereas the Big model obtains 33.6 and 37.9 BLEU, respectively. This represents an increase of 2.7 BLEU in both the dev and the test sets. As in the French-English case, the use of synthetic backtranslations can be explored in order to improve these results. As in the French-German, the BLEU score achieved of 26.2 on VL is lower than that on the WMT task.

4.7 Italian-English

For the Italian-English systems, we have collected a series of public datasets from a variety of domains such as: medical (EMEA) and institutional documents (ECB, Europarl and JRC-Acquis), book translations (EUbookshop) and Wikipedia [65]. The statistics of these datasets are shown in Table 23.

Table 23: Statistics of the data sets used to train the Italian-English MT systems.

Corpus	Sentences(K)	Words(M)		Vocabulary(K)	
		It	En	It	En
ECB	193.0	5.8	5.5	77.7	62.2
EMEA	1081.1	13.4	12.1	153.7	130.3
EUbookshop	6490.0	147.4	144.6	2515.8	2332.6
Europarl-v7	1944.9	49.0	50.7	492.3	380.1
JRC-Acquis	811.0	15.4	15.5	248.4	217.6
Wikipedia	957.0	19.2	20.6	1520.1	1530.0

In this case, we have trained two Transformer Base models, one trained with 1 GPU and the other with 3 GPU, with a maximum sequence length of 100.

The WMT competition has not been held for the Italian-English pair. As such, we must look elsewhere to find reliable test sets. In this case, we have used the tests sets from IWSLT17 [66], another international MT competition. We used the provided dev and test sets for Italian-English. Table 24 shows the results obtained by the Italian-English MT systems.

Table 24: Evaluation results of the Italian-English MT systems on the IWSLT task.

System	dev BLEU	test BLEU
Transformer Base	25.1	25.4
Transformer Base, 3 GPU	25.1	25.9

The Transformer Base achieves 25.1 BLEU in the dev set, and 25.4 BLEU in the test set, and the 3-GPU version improves 0.5 BLEU in the test set. Although not very significant, we also observe performance differences due to different batch sizes. As the amount of available data is lower in this language pair when compared with others, we believe the use of additional data in the way of backtranslations would be specially beneficial in this case. Additionally, bigger architectures have not been assessed yet. Therefore, future work goes in this line to improve model performance.

4.8 English-Italian

The resources chosen to build the English-Italian system are the same as those of the Italian-English MT system, as the language pair involved is the same. The details are shown in Table 23.

Following the Italian-English setup, we train both, a Transformer Base with 1 GPU and with 3 GPUs. Table 25 shows the results obtained by the English-Italian MT systems.

Table 25: Evaluation results of the English-Italian MT systems on the IWSLT task.

System	dev BLEU	test BLEU
Transformer Base	21.4	21.4
Transformer Base, 3 GPU	23.7	23.3

The 1 GPU Transformer Base models obtains 21.4 BLEU in both the dev and the test set. The 3-GPU Transformer obtains an improvement of 2.3 and 1.9 BLEU, respectively. As the only difference between these two models is the batch size, this results prove that the choice of batch size is critical for Transformer models. As in the Italian-English system, significant improvements can be achieved by increasing the amount of training data.

4.9 Slovenian-English

For the Slovenian-English system, we have collected a series of public datasets from a variety of domains such as: talks (VL.NET, WIT and TEDx), institutional documents (DGT, DGT-TM and Europarl), medical (EMEA), book translations (EUbookshop) and broadcast TV (EUTV and OpenSubtitles) [65]. The statistics of these datasets are shown in Table 26.

As for the Italian pairs mentioned above, the WMT competition has not been held for the English-Slovenian pair. So, as in the Italian pairs, we revert to the IWSLT dev/test sets released in the 2012, 2013 and 2014 editions as out-domain task. Parameter tuning was performed on the VL dev set. As in previous systems, a Transformer Base model with 1 GPU and maximum sequence length of 75 words was trained.

Table 27 shows the results obtained by the Slovenian-English MT system on IWSLT sets. As observed, all BLEU figures are above 30. A notable exception is the VL test set that achieves 26.4 BLEU points. The explanation why this VL test set is specially difficult, it is because the reference manual translations into Slovenian provided by professional translators were not so literal as in the WMT/IWSLT test sets.

Table 26: Statistics of the data sets used to train the Slovenian-English MT systems.

Corpus	Sentences(K)	Words(M)		Vocabulary(K)	
		Sl	En	Sl	En
VL.NET	17.1	0.4	0.4	33.3	15.1
WIT	17.0	0.2	0.3	32.4	16.7
TEDx	53.8	0.4	0.5	41.0	20.6
DGT	3099.3	50.7	55.8	422.0	308.3
DGT-TM	2496.7	40.6	44.7	381.0	274.7
Europarl	617.0	14.0	16.1	144.4	66.1
EMEA	1033.2	13.3	13.1	95.5	60.0
EUbookshop	391.4	8.2	8.9	256.4	182.9
EUTV	181.3	1.7	2.0	62.1	29.1
OpenSubtitles	19636.1	127.6	167.9	1011.1	574.5
Total	27528.6	256.9	309.4	1497.6	970.0

Table 27: Evaluation results of the Slovenian-English MT system on the IWSLT task.

	dev2012	tst2012	test2013	tst2014
Transformer Base	32.1	31.4	34.3	32.1

4.10 English-Slovenian

The English-Slovenian system is trained and evaluated on the same data sets as its counterpart Slovenian-English and the same MT architecture: Transformer Base model trained on 1 GPU and maximum sequence length of 75 words.

Table 28 shows the results obtained by the English-Slovenian MT system on VL and IWSLT sets. As observed, BLEU scores are similar across corpora in most cases although lower than Slovenian-English. This is explained by the complexity of translating into Slovenian.

Table 28: Evaluation results of the English-Slovenian MT system on the WMT task.

	IWSLT			
	dev2012	tst2012	test2013	tst2014
Transformer Base	27.8	25.8	29.4	27.8

As in the Slovenian-English pair, the BLEU score achieved 22.9 is lower than that of the IWSLT tasks.

4.11 German-French

For the German-French system, we have used the dataset available for the WMT19 German-French news translation task: commoncrawl, europarl, news-commentary and paracrawl. The statistics of these datasets are shown in Table 29.

We extracted 1M sentence pairs from the paracrawl corpus using Cross-Entropy filtering. We also used 10M backtranslations produced with the French-German Transformer Base system. Using this data, we have trained two systems, a Transformer Base model that was trained only with the original data, and a Transformer Big model, trained with the additional filtered data and backtranslation.

We have used the provided WMT19 evaluation sets for this task. We split the dev set provided (from EU elections) into two sets, dev1 and dev2, using the former as dev set, and the later as internal

Table 29: Statistics of the WMT German-French dataset.

Corpus	Sentences(K)	Words(M)		Vocabulary(K)	
		De	Fr	De	Fr
Common Crawl	622.3	12.2	14.0	932.2	676.8
Europarl-v7	1726.4	41.0	46.0	616.7	388.6
News Commentary v14	263.2	5.9	6.7	281.8	207.5
ParaCrawl v3	7222.6	99.7	110.7	4022.4	3083.1

test set. We also report results with the competition's official test set, newstest2019. The results of these systems are shown in Table 30.

Table 30: Evaluation results of the German-French MT systems on the WMT task.

System	dev2	newstest2019
Transformer Base	31.1	32.1
Transformer Big, 4 GPU + Backtrans.	33.3	34.4

The Transformer Base obtains 31.1 BLEU on dev2, and 32.1 BLEU in newstest2019. The Transformer Big improves these results by 2.2 and 2.3 BLEU, respectively. This improvement is achieved thanks to the use of the Transformer Big architecture as well as additional synthetic data.

The BLEU score on the VL task was 18.6, that is among the lowest BLEU scores in the VL task. We believe this is explained by the fact that the German and French sentences are not direct translations of each other, but translations from a common source in English.

4.12 French-German

The resources chosen to build the French-German system are the same as those of the German-French MT system, as the language pair involved is the same. The details are shown in Table 29.

Following the same setup as in the German-French, we have trained both, a Transformer Base and a Transformer Big system, with the later system using the paracrawl filtered data as well as 18M backtranslations produced by the German-French Transformer Base model. The results of these systems, evaluated on the WMT19 sets, are shown in Table 31.

Table 31: Evaluation results of the French-German MT systems on the WMT task.

System	dev2	newstest2019
Transformer Base	22.8	25.7
Transformer Big, 4 GPU + Backtrans.	24.9	26.9

The Transformer Base model obtains 22.8 BLEU in the dev2 set, and 25.7 BLEU in newstest2019. The Transformer Big obtains improvements of 2.1 and 1.2 BLEU, respectively. Even though the German-French and French-German datasets are the same, these results show how translating into German is a more difficult task.

Similarly to the German-French pair, the BLEU score is 17.2 on the VL task, as mentioned before we believe that the main reason for these results is the procedure followed to artificially generate French-German parallel sentence pairs from English.

4.13 Portuguese-Spanish

We have developed NMT systems for Portuguese-Spanish using the data setup of the WMT19 Similar Language Translation Task. We split the provided development data into two sets, and used one as dev set and the other as test set. The corpora available for this task is shown in Table 32. We found out that there is a significant domain mismatch between the available training data and the test data. Where as the former mostly consists in institutional recordings and documents, the later is an specific website translation task.

Table 32: Statistics of the data sets used to train the Portuguese-Spanish MT systems.

Corpus	Sent.(K)	Words(M)		Vocab.(K)	
		Es	Pt	Es	Pt
JCR Acquis	1650	42	40	264	264
Europarl v9	1812	53	52	177	156
News Commentary v14	48	1	1	49	47
Wiki Titles v1	621	1	1	292	295

Due to the aforementioned domain mismatch, we focused our efforts in using domain adaptation techniques, as we hypothesized that this could provide greater improvements than other techniques such as backtranslations or higher model capacity. In order to carry out domain adaptation, we used the fine-tuning schema, where we trained an initial model with a large out-domain data until convergence, and then we continue training this model on a small in-domain corpus. We simply selected a small subset of the dev set as in-domain training data. We tested this technique with a Transformer Base model trained with 4 GPU. The results are shown in Table 33.

Table 33: Evaluation results of the Portuguese-Spanish MT systems.

	BLEU	
System	test	test-hidden
Transformer Base, 4GPU	57.4	51.9
+ fine-tuned	72.4	66.6

The baseline Transformer obtains 57.4 BLEU in the test set and 51.9 BLEU in the competition' test set. The fine-tuned version obtains substantial improvements over the baseline, with an increase of 15.0 and 14.7 BLEU, respectively. This confirms our initial hypothesis about domain mismatch. This fine-tuned model was submitted to the competition, and it outperformed all other participants by a significant margin, with an improvement of 6.7 BLEU over the second-best system. It was not possible to carry out a comparison with Google because the reference translations for the competition's test set have not been published.

4.14 Spanish-Portuguese

For the Spanish-Portuguese system, we also followed the setup of the WMT19 Similar Language task, summarized in Table 32. In the same way as the Portuguese-Spanish case, we trained a 4 GPU Transformer Base model, and used the fine-tuning technique in order to carry out domain-adaptation. The results are shown in Table 34.

Following the trend of the Portuguese-Spanish system, the fine-tuning version obtains significant improvements over the Transformer baseline. In this case, the fine-tuned version obtains an improvement of 19.4 BLEU in the test set, and a 19.2 BLEU in the competition's test set. The fine-tuned

Table 34: Evaluation results of the Spanish-Portuguese MT systems.

	BLEU	
System	test	test-hidden
Transformer	51.3	45.5
+ fine-tuned	70.7	64.7

model outperformed all other submissions to the Spanish-Portuguese task, with an improvement of 2.6 BLEU over the second-best system.

4.15 Comparative results with Google Translate

In this section, we report comparative results between the best X5gon systems in each language pair and Google Translate, the main reference online translation service that is publicly available. To this end, the VL and WMT/IWSLT test sets previously described were automatically translated using Google Translate and BLEU scores were computed on the translations returned. In case there were several WMT/IWSLT test sets available, that with the highest BLEU score is reported.

Table 35 shows the BLEU scores described above. As observed, Google Translate achieves slightly better or similar results in most pairs involving languages in which public data resources are abundant on the Internet. However, BLEU scores were lower in Slovenian for which limited data resources exist and the stopping training criteria was based on an in-domain dev set. It is also worth noting the significant better results in the Portuguese pairs in which fine-tuning was performed using an in-domain training and dev sets.

Table 35: Comparative results between X5gon and Google Translate.

	De-En		En-De		Es-En		En-Es		Fr-En		En-Fr	
	VL	WMT	VL	WMT	VL	WMT	VL	WMT	VL	WMT	VL	WMT
X5gon	**27.0**	**48.0**	21.5	45.7	36.4	32.3	39.4	32.2	29.0	36.8	26.2	37.9
Google	25.7	43.9	**24.7**	**47.0**	**37.8**	**34.4**	**41.3**	**35.3**	**30.3**	**38.6**	**29.4**	**40.4**

	It-En	En-It	Sl-En		En-Sl		De-Fr		Fr-De		Pt-Es	Es-Pt
	IWSLT	IWSLT	VL	IWSLT	VL	IWSLT	VL	WMT	VL	WMT	WMT	WMT
X5gon	25.9	23.3	**26.4**	**34.3**	**22.9**	**29.4**	18.6	**34.4**	17.2	**26.9**	**72.4**	**70.7**
Google	**35.7**	**32.1**	15.0	29.2	16.5	23.6	**19.6**	32.2	**18.6**	26.6	47.6	43.4

4.16 Conclusions and future work

We have presented the performance of the MT systems in a series of language pairs on in-domain (VL) and out-domain (WMT/IWSLT) test sets. As observed in Figure 1(b), Portuguese, German and French systems tuned and evaluated on out-domain data exhibit BLEU scores above 35 that can be considered fairly accurate MT systems to be deployed in a production real-world environment. Just below 35 BLEU points, we have the Spanish language pairs, Slovene-English and German-French assessed on out-domain test sets, that would require little effort to achieve accurate enough performance applying some of the advanced techniques studied in the German pairs. Below 30 BLEU points, we find most in-domain evaluations on Videolectures.NET in which we will apply a bigger batch size and/or in-domain fine-tuning to fill the domain-mismatch BLEU gap, as we did for Portuguese. In the specific case of the Italian pairs, in addition to what mentioned above, we will need to increase the volume of training data to boost BLEU scores.

References

[1] D5.1: First report on piloting. Technical report, X5gon project, M12, 2018.

[2] GSC-TUDa: German Speech Corpus by Technische Universität Darmstadt. `https://www.lt.informatik.tu-darmstadt.de/de/data/open-acoustic-models`.

[3] Wikipedia. `https://www.wikipedia.org/`.

[4] Europarl Corpus: European Parliament Proceedings Parallel Corpus v7. `http://www.statmt.org/europarl/`.

[5] commoncrawl 2014. `http://commoncrawl.org/`.

[6] News Crawl corpus (from WMT workshop) 2015. `http://www.statmt.org/wmt15/translation-task.html`.

[7] REUTERS: Reuters Corpora (RCV1, RCV2, TRC2). `http://trec.nist.gov/data/reuters/reuters.html`

[8] Tatoeba. `https://tatoeba.org/eng/downloads`.

[9] WEBCELEX: The CELEX Lexical Database (English, Dutch and German word features). `http://celex.mpi.nl/`.

[10] M.J.F. Gales. Maximum likelihood linear transformations for hmm-based speech recognition. *Computer Speech and Language*, 12(2):75 – 98, 1998.

[11] Geoffrey Hinton, Li Deng, Dong Yu, Abdel rahman Mohamed, Navdeep Jaitly, Andrew Senior, Vincent Vanhoucke, Patrick Nguyen, Tara Sainath George Dahl, and Brian Kingsbury. Deep neural networks for acoustic modeling in speech recognition. *IEEE Signal Processing Magazine*, 29(6):82–97, November 2012.

[12] G. Stemmer, F. Brugnara, and D. Giuliani. Adaptive training using simple target models. In *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, volume 1, pages 997 – 1000, 2005.

[13] M.A. del Agua, A. Giménez, N. Serrano, J. Andrés-Ferrer, J. Civera, A. Sanchis, and A. Juan. The translectures-upv toolkit. In JuanLuis Navarro Mesa, Alfonso Ortega, António Teixeira, Eduardo Hernández Pérez, Pedro Quintana Morales, Antonio Ravelo García, Iván Guerra Moreno, and DoroteoT. Toledano, editors, *Advances in Speech and Language Technologies for Iberian Languages*, volume 8854 of *Lecture Notes in Computer Science*, pages 269–278. Springer International Publishing, 2014.

[14] TensorFlow. `https://www.tensorflow.org/`.

[15] TED. `https://www.ted.com/talks`.

[16] The RNNLM Toolkit . `http://www.rnnlm.org/`.

[17] Sequitur G2P. `http://www-i6.informatik.rwth-aachen.de/web/Software/g2p.html`.

[18] transLectures. `http://www.translectures.eu/web`.

[19] Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. Librispeech: an ASR corpus based on public domain audio books. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5206–5210, 2015.

[20] Common Voice. https://voice.mozilla.org/en.

[21] Javier Jorge, Adrià Martínez-Villaronga, Pavel Golik, Adrià Giménez, Joan Albert Silvestre-Cerdà, Patrick Doetsch, Vicent Andreu Císcar, Hermann Ney, Alfons Juan, and Albert Sanchis. MLLP-UPV and RWTH Aachen Spanish ASR Systems for the IberSpeech-RTVE 2018 Speech-to-Text Transcription Challenge. In *Proc. of IberSPEECH 2018: 10th Jornadas en Tecnologías del Habla and 6th Iberian SLTech Workshop*, pages 257–261, Barcelona (Spain), 2018.

[22] François Hernandez, Vincent Nguyen, Sahar Ghannay, Natalia Tomashenko, and Yannick Estève. TED-LIUM 3: twice as much data and corpus repartition for experiments on speaker adaptation. In *International Conference on Speech and Computer*, pages 198–208. Springer, 2018.

[23] The Spoken Wikipedia Corpora. http://nats.gitlab.io/swc/.

[24] VideoLectures.NET. http://videolectures.net/.

[25] VoxForge. http://www.voxforge.org/.

[26] AMI Corpus. http://groups.inf.ed.ac.uk/ami/corpus/.

[27] EPPS Speech corpus at ELRA. http://catalog.elra.info/product_info.php?products_id=1035.

[28] The ELFA Corpus. http://www.uta.fi/ltl/en/english/research/projects/elfa/corpus.html.

[29] CSTR VCTK Corpus. http://homepages.inf.ed.ac.uk/jyamagis/page3/page58/page58.html.

[30] poliMedia. https://media.upv.es/#/catalog.

[31] Ondřej Bojar, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Philipp Koehn, and Christof Monz. Findings of the 2018 conference on machine translation (WMT18). In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 272–303, Belgium, Brussels, October 2018. Association for Computational Linguistics.

[32] Google Books count (version 2, tagged 20120701). http://storage.googleapis.com/books/ngrams/books/

[33] 10^9. http://www.statmt.org/wmt10/training-giga-fren.tar.

[34] Andreas Eisele and Yu Chen. Multiun: A multilingual corpus from united nation documents. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis, Mike Rosner, and Daniel Tapias, editors, *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta, may 2010. European Language Resources Association (ELRA).

[35] HyperArticles en ligne (HAL). http://hal.archives-ouvertes.fr/.

[36] DGT-Translation Memory. https://ec.europa.eu/jrc/en/language-technologies/dgt-translation-m

[37] News Commentary (bilingual). http://www.statmt.org/wmt13/training-parallel-nc-v8.tgz.

[38] News Commentary (Monolingual). http://www.statmt.org/wmt13/training-monolingual-nc-v8.tgz.

[39] Mauro Cettolo, Christian Girardi, and Marcello Federico. WIT^3: Web Inventory of Transcribed and Translated Talks. In *Proceedings of the 16th Conference of the European Association for Machine Translation (EAMT)*, pages 261–268, Trento, Italy, May 2012.

[40] Patrik Lambert, Jean Senellart, Laurent Romary, Holger Schwenk, Florian Zipser, Patrice Lopez, and Frédéric Blain. Collaborative machine translation service for scientific texts. In *EACL*, pages 11–15, 2012.

[41] EuroParl TV. `www.europarltv.europa.eu`.

[42] Albert Zeyer, Patrick Doetsch, Paul Voigtlaender, Ralf Schlüter, and Hermann Ney. A comprehensive study of deep bidirectional lstm rnns for acoustic modeling in speech recognition. In *IEEE International Conference on Acoustics, Speech, and Signal Processing*, pages 2462–2466, New Orleans, LA, USA, March 2017.

[43] Xi Chen, Xin Liu, Y. Qian, Mark J. F. Gales, and Philip C. Woodland. Cued-rnnlm — an open-source toolkit for efficient training and evaluation of recurrent neural network language models. *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6000–6004, 2016.

[44] Javier Jorge, Adrià Giménez, Javier Iranzo-Sánchez, Jorge Civera, Albert Sanchis, and Alfons Juan. Real-time one-pass decoder for speech recognition using lstm language models. In *Proc. of the 20th Annual Conf. of the ISCA (Interspeech 2019)*, Graz (Austria), 2019. in press.

[45] Joan Albert Silvestre-Cerdà, Adrià Giménez, Jesús Andrés-Ferrer, Jorge Civera, and Alfons Juan. Albayzin Evaluation: The PRHLT-UPV Audio Segmentation System. In *Proceedings of IberSPEECH 2012*, pages 596–600, Madrid (Spain), 2012.

[46] M. Cettolo, J. Niehues, S.Stüker, L. Bentivogli, R. Cattoni, and M. Federico. The IWSLT 2015 Evaluation Campaign. In *Proc. of 12th Intl. Workshop on Spoken Language Translation (IWSLT 2015)*, pages 2–10, Da Nang (Vietnam), 2015.

[47] A. Martínez-Villaronga, M.A. del Agua, J. Andrés-Ferrer, and A. Juan. Language model adaptation for video lectures transcription. In *Proc. of the IEEE Intl. Conf. on Acoustics, Speech and Signal Processing ICASSP 2013*, pages 8450–8454, Vancouver (Canada), 2013.

[48] Ana Zwitter Vitez, Jana Zemljarič Miklavčič, Simon Krek, Marko Stabej, and Tomaž Erjavec. Spoken corpus gos 1.0, 2013. Slovenian language resource repository CLARIN.SI.

[49] Darinka Verdonik, Tomaž Potočnik, Mirjam Sepesy Maučec, Tomaž Erjavec, Simona Majhenič, and Andrej Žgank. Spoken corpus gos VideoLectures 4.0 (transcription), 2019. Slovenian language resource repository CLARIN.SI.

[50] UPVLC, XEROX, JSI-K4A, RWTH, and EML. D3.1.3: Final report on massive adaptation. Technical report, transLectures, 2014.

[51] Nataša Logar, Tomaž Erjavec, Simon Krek, Miha Grčar, and Peter Holozan. Written corpus ccGigafida 1.0, 2013. Slovenian language resource repository CLARIN.SI.

[52] Nikola Ljubešić and Tomaž Erjavec. hrwac and slwac: Compiling web corpora for croatian and slovene. In Ivan Habernal and Václav Matousek, editors, *Text, Speech and Dialogue - 14th International Conference, TSD 2011, Pilsen, Czech Republic, September 1-5, 2011. Proceedings*, Lecture Notes in Computer Science, pages 395–402. Springer, 2011.

[53] Andrej Zgank, Mirjam Sepesy Maucec, and Darinka Verdonik. The si tedx-um speech database: a new slovenian spoken language resource. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Sara Goggi, Marko Grobelnik, Bente Maegaard, Joseph Mariani,

Helene Mazo, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Paris, France, may 2016. European Language Resources Association (ELRA).

[54] Mauro Cettolo, Christian Girardi, and Marcello Federico. Wit^3: Web inventory of transcribed and translated talks. In *Proceedings of the 16th Conference of the European Association for Machine Translation (EAMT)*, pages 261–268, Trento, Italy, May 2012.

[55] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, 4-9 December 2017, Long Beach, CA, USA*, pages 6000–6010, 2017.

[56] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, pages 311–318, Stroudsburg, PA, USA, 2002. Association for Computational Linguistics.

[57] Ondrej Bojar, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Philipp Koehn, and Christof Monz. Findings of the 2018 conference on machine translation (WMT18). In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers, WMT 2018, Belgium, Brussels, October 31 - November 1, 2018*, pages 272–303, 2018.

[58] News Commentary v13. `http://data.statmt.org/wmt18/translation-task/training-parallel-nc-v13`

[59] Roberts Rozis and Raivis Skadiņš. Tilde MODEL - multilingual open data for EU languages. In *Proceedings of the 21st Nordic Conference on Computational Linguistics*, pages 263–265, Gothenburg, Sweden, May 2017. Association for Computational Linguistics.

[60] ParaCrawl. `https://paracrawl.eu/`.

[61] Javier Iranzo-Sánchez, Pau Baquero-Arnal, Gonçal V. Garcés Díaz-Munío, Adria A. Martinez-Villaronga, Jorge Civera, and Alfons Juan. The MLLP-UPV german-english machine translation system for WMT18. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers, WMT 2018, Belgium, Brussels, October 31 - November 1, 2018*, pages 418–424, 2018.

[62] Rico Sennrich, Barry Haddow, and Alexandra Birch. Improving neural machine translation models with monolingual data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016, August 7-12, 2016, Berlin, Germany, Volume 1: Long Papers*, 2016.

[63] Marcin Junczys-Dowmunt. Dual conditional cross-entropy filtering of noisy parallel corpora. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers, WMT 2018, Belgium, Brussels, October 31 - November 1, 2018*, pages 888–895, 2018.

[64] Sergey Edunov, Myle Ott, Michael Auli, and David Grangier. Understanding back-translation at scale. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, pages 489–500, 2018.

[65] Jörg Tiedemann. Parallel data, tools and interfaces in OPUS. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC-2012)*, pages 2214–2218, Istanbul, Turkey, May 2012. European Languages Resources Association (ELRA).

[66] Mauro Cettolo, Marcello Federico, Luisa Bentivogli, Niehues Jan, Stüker Sebastian, Sudoh Katsuitho, Yoshino Koichiro, and Federmann Christian. Overview of the iwslt 2017 evaluation campaign. In *International Workshop on Spoken Language Translation*, pages 2–14, 2017.

[67] Davor Orlic. D7.1: Website. Technical report, X5gon project, M1, 2018.

[68] Erik Novak. D2.1: Requirements & Architecture Report. Technical report, X5gon project, M6, 2018.

[69] Stefan Kreitmayer. D4.1: Initial prototype of user modelling architecture. Technical report, X5gon project, M6, 2018.

[70] D1.1: Quality assurance models. Technical report, X5gon project, M12, 2018.

[71] D1.2: Report on selected and evaluated quality assurance models. Technical report, X5gon project, M12, 2018.

[72] D3.1: Learning Analytic Engine 2.0. Technical report, X5gon project, M12, 2018.

[73] D6.1: Report of the OER network model and interface design evaluation. Technical report, X5gon project, M12, 2018.

[74] D7.2: First real-world and online community engagement plan. Technical report, X5gon project, M12, 2018.

[75] D8.1: Detailed market analysis. Technical report, X5gon project, M12, 2018.

[76] D9.1: Ethical Data. Management and Data. Management Pla: year 1. Technical report, X5gon project, M12, 2018.

[77] D9.4: First year report. Technical report, X5gon project, M12, 2018.

[78] EMMA. http://project.europeanmoocs.eu/.

[79] poliMedia. https://politrans.upv.es/.

[80] Review Report. Technical report, X5gon project, November 2018 (M15).

[81] Jorge Civera and Alfons Juan. T36: Final report. Technical report, UPV, 2014.

[82] The MLLP Transcription and Translation Platform (MLLP-TTP). https://ttp.mllp.upv.es.

[83] Juan Daniel Valor Miró, Pau Baquero-Arnal, Jorge Civera, Carlos Turró, and Alfons Juan. Multilingual Videos for MOOCs and OER. *Educational Technology & Society*, 21(2):1–12, 2018.

[84] UPVLC. D2.3.2: Report on final transcription and translation models. Technical report, EMMA, 2015.

[85] Rico Sennrich, Barry Haddow, and Alexandra Birch. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016, August 7-12, 2016, Berlin, Germany, Volume 1: Long Papers*, 2016.