Overview of the IWSLT 2017 Evaluation Campaign

M. Cettolo[1] *M. Federico*[1] *L. Bentivogli*[1] *J. Niehues*[2]
S. Stüker[2] *K. Sudoh*[3] *K. Yoshino*[3] *C. Federmann*[4]

[1] FBK - Trento, Italy
[2] KIT - Karlsruhe, Germany
[3] NAIST - Nara, Japan
[4] Microsoft AI+Research - Redmond, WA, USA

Abstract

The IWSLT 2017 evaluation campaign has organised three tasks. The Multilingual task, which is about training machine translation systems handling many-to-many language directions, including so-called zero-shot directions. The Dialogue task, which calls for the integration of context information in machine translation, in order to resolve anaphoric references that typically occur in human-human dialogue turns. And, finally, the Lecture task, which offers the challenge of automatically transcribing and translating real-life university lectures. Following the tradition of these reports, we will described all tasks in detail and present the results of all runs submitted by their participants.

1. Introduction

Spoken language translation (SLT) is the subfield of machine translation (MT) that deals with the translation of spoken language. Spoken language, besides differing from written language from a linguistic point of view [1], also implies that it is processed under form of a transcript, either manually created and cleaned or generated via automatic speech recognition (ASR) and thus possibly noisy.

Since 2004, the International Workshop on Spoken Language Translation has been organizing a yearly evaluation campaign in conjunction with a scientific workshop. The main purpose of the evaluation campaigns is to offer to researchers working in the fields of MT and ASR challenging tasks to work on, as well as providing for them a venue where to present, compare and discuss their results. Moreover, in order to offer a friendly environment for scientific exchange, the spirit of our evaluation has never been competitive, but rather collaborative.

The tasks offered during the last 13 years have followed the trend and progress in the field of MT and ASR. In the first years, SLT tasks focused on restricted domains, with low language complexity. Then, following the steady rise of statistical methods and computing power, less restricted and more data intensive tasks were progressively introduced, up to the translation of TED Talks and university lectures. However, in order to keep the participation barrier low, IWSLT has also always offered at the same time tasks that were affordable to small teams or even students with limited access to computing resources. Another distinctive feature of IWSLT is the variety of translation directions covered over the years, which include many American, European and Asian languages.

We believe that scientific communication is greatly facilitated when all experimental conditions are set in advance and shared by everyone. This is the reason why, since the begin, IWSLT has organized shared tasks in which all the training data, experimental conditions and evaluation metrics were set and provided in advance.

This year, the IWSLT evaluation campaign has focused on three tasks, which address rather different and orthogonal open issues in MT, in general, and spoken language translation, in particular. The Multilingual task investigates the possibility of machines to simultaneously learn to translate across multiple languages, given parallel data (TED Talks) that only partially covers the tested translation directions. The Dialogues task targets instead the challenge for MT to consider the context of the input (utterance transcript) that has to be translated, in order to resolve the translation of pronouns and other empty categories. Finally, the Lecture task addresses the challenge of automatically transcribing and translating real-life university lectures, in contrast of staged and well-rehearsed talks, such as the TED Talks.

The following sections describe in great details each task, including the benchmark that has been developed around it and the outcome of the evaluation. One specific section will be devoted to report on the manual evaluation that was car-

ried out for the Multilingual task. An appendix concludes this report, which contains all the tables with the results of all the submitted runs.
Finally, this year we have witnessed, unfortunately, a significant drop in the number of participants to the evaluation campaign (see Table 1). For this reason, part of the open discussion that will take place at the workshop will regard this issue. Our aim will be to understand if the lack of participation has a contingent nature or expresses a shift of interest in the community. In either cases, as organizers, we will see if and how we can find better ways to serve the community.

2. Multilingual Task

2.1. Definition

The introduction of translation of TED talks in IWSLT evaluation campaigns dates back to 2010. The task continues to receive attention by the research community because it is challenging but at the same time manageable. In fact, besides being a realistic exercise, the variety of topics dealt with in TED talks can be considered unlimited, which is an interesting research issue in itself. On the other hand, the truly "in-domain" training data, that is the set of transcriptions and translations of TED talks only, amount to just few million words per side, making the training/adaptation of even neural engines reasonably fast.

With the aim of keeping the task interesting and to follow current trends in research and industry, this year we proposed the multilingual translation between any pair of languages from {Dutch, English, German, Italian, Romanian} by means of an engine trained with either only in-domain data (*small data condition*) or a long list of permissible resources (*large data condition*). In addition, within the small condition, we proposed the *zero-shot* translation for the pairs Dutch-German and Italian-Romanian, in both directions. Zero-shot means to translate with a multilingual engine between language pairs that have never seen in this combination during training. In the specific, the zero-shot engine could be trained on the in-domain training data of all the other 16 pairs, but not of those four pairs. Training data synthesis from the 16 pairs and pivoting were explicitly forbidden, in order to force the adoption of methods that deal with the problem instead of getting around it. The zero-shot paired languages are from the same family (West-Germanic and Romance, respectively) in the hope that they can somehow leverage from their common origin.

A set of unofficial standard bilingual tasks between English from one side and {Arabic, Chinese, French, German, Japanese, Korean} on the other were proposed as well to keep continuity with past editions.

2.2. Data

In-domain training, development and evaluation sets were supplied through the website of the WIT[3] project [9], while out-of-domain training data were linked in the workshop's website. With respect to edition 2016 of the evaluation campaign, some of the talks added to the TED repository during the last year have been used to define the evaluation sets (tst2017), while the remaining new talks have been included in the training sets.

Two development sets (dev2010 and tst2010) are either the same of past editions - when available - or have been built upon the same talks - for pairs never proposed in the past.

Table 2 provides statistics on in-domain texts supplied for training, development and evaluation purposes, averaged on the 20 language pairs.

Concerning the unofficial bilingual task, besides the tst2017 evaluation set, we asked to translate the progressive tst2016 test set as well.

2.3. Evaluation

Participants had to provide MT outputs of the test sets in NIST XML format. Outputs had to be case-sensitive, detokenized and punctuated. The quality of translations was measured both automatically, against human translations created by the TED open translation project, and via human evaluation (Section 5). Case sensitive automatic scores were calculated with the three automatic standard metrics BLEU, NIST, and TER, as implemented in mteval-v13a.pl[1] and tercom-0.7.25[2], by calling:

- `mteval-v13a.pl -c`
- `java -Dfile.encoding=UTF8 -jar tercom.7.25.jar -N -s`

Detokenized texts were used, since the two scoring scripts apply their own internal tokenizers.

In order to allow participants to evaluate their progresses automatically and under identical conditions, an evaluation server was set up. Participants could submit the translation of any development set to either a REST Webservice or through a GUI on the web, receiving as output BLEU, NIST and TER scores computed as described above.

The evaluation server was utilized by the organizers for the automatic evaluation of the official submissions. After the evaluation period, the

[1] http://www.itl.nist.gov/iad/mig/tests/mt/2009/
[2] 5http://www.cs.umd.edu/ snover/tercom/

Table 1: List of Participants

FBK	Fondazione Bruno Kessler, Italy [2]
GTCT	Global Tone Communication Technology Co. Ltd, China[3]
KIT	Karlsruhe Institute of Technology, Germany [4]
KYOTO	Kyoto University, Japan [5]
RWTH	Rheinisch-Westfälische Technische Hochschule, Germany [6]
UEDIN	University of Edinburgh, United Kingdom [7]
UDSDFKI	Universität des Saarlandes and Deutsche Forschungszentrum für Künstliche Intelligenz, Germany [8]

Table 2: Average size of bilingual resources made available for the 20 language pairs of the multilingual task.

data set	sent	tokens source	tokens target	talks
train	160k	3.99M	3.99M	1749
dev2010	940	18,8k	18,8k	8
tst2010	1,660	30,0k	30,0k	11
tst2017	1,146	19,8k	19,8k	10

evaluation of test sets was allowed to all participants as well.

2.4. Submissions

We received 9 primary multilingual submissions from 5 different sites, distributed according to training conditions as follows: 4 on small-data, 4 on zero-shot and 1 on large-data; in addition, 3 small-data, 2 zero-shot and 1 large-data contrastive runs were submitted. One out of those five participants also sent a bilingual run on Chinese-English, while two other participants provided their runs on German-English bilingual task.

The total number of test sets evaluated for the multilingual task was then 300 (180 primary, 120 contrastive), while as far as the bilingual tasks are concerned, 12 translations were scored.

2.5. Automatic results

The automatic scores computed on the 2017 official test set for each participant are shown in Appendix A. The two uppermost tables concern the four zero-shot language pairs, where scores of all multilingual submissions are provided.

Table 3 reports the automatic scores of the 9 primary multilingual submissions averaged on the four directions involving the zero-shot condition. Despite being questionable, the average operation allows to synthesize some general outcomes in a easier way than looking at the many tables of the appendix:

- as proved by KYOTO, zero-shot systems ("ML ZS") can well compete with those trained including data of the language pairs they are tested on ("ML SD")
- also other labs were able to develop zero-shot systems reasonably good with respect to their best systems, endorsing the general feasibility of zero-shot translation
- KIT, the only lab that submitted runs for both small- and large-data conditions, was able to reach the highest MT quality by using more data for training, but not by far. Such performance proximity could be due to multilinguality, which allows the weaker condition (SD) to handle sparsity, problem that does not affect too much the LD engine. In other words, multilinguality seems to represent an effective solution to data sparsity, alternative to the use of large out-of-domain data sets.

Table 3: Automatic scores of the primary multilingual submissions averaged on the four zero-shot language pairs.

system	**cond.**	BLEU	NIST	TER
FBK	ML SD	19.54	5.432	62.81
	ML ZS	17.26	5.077	65.29
GTCT	ML ZS	19.40	5.343	63.27
KIT	ML SD	20.97	5.716	60.38
	ML LD	21.13	5.765	59.77
KYOTO	ML SD	20.60	5.621	61.54
	ML ZS	20.55	5.573	61.84
UDSDFKI	ML SD	19.06	5.342	64.26
	ML ZS	17.10	5.088	65.81

Table 4 reports the automatic scores of the 9 primary multilingual submissions averaged on the 16 directions other than the zero-shot. For these directions, the ML ZS systems are not at all "zero-shot" systems, but simply multilingual systems trained on parallel data for 16 pairs, including that which they are tested on. Therefore, the table compares multilingual systems trained

on either 20 or 16 pairs. In one case (FBK) the ML SD system is better than the ML ZS, in another (KYOTO) it is the opposite, while in the third case (UDSDFKI) they perform equally; no general conclusion can be drawn for now but the issue deserves further investigation.

Table 4: Automatic scores of the primary multilingual submissions averaged on the 16 non zero-shot language pairs.

system	cond.	BLEU	NIST	TER
FBK	ML SD	22.31	5.818	59.89
	ML ZS	21.89	5.760	60.36
GTCT	ML ZS	24.46	6.112	57.61
KIT	ML SD	24.07	6.139	57.12
	ML LD	24.42	6.191	56.56
KYOTO	ML SD	23.73	6.059	58.00
	ML ZS	24.10	6.083	57.78
UDSDFKI	ML SD	21.69	5.764	60.75
	ML ZS	21.63	5.749	60.89

3. Dialogue Task

3.1. Definition

Despite the recent advances of machine translation technologies, their effectiveness has not been investigated well by highly context-dependent situations such as dialogues. One typical problem in the translation of dialogues is the existence of empty categories [10], especially in pro-drop source languages such as Chinese, Japanese, and Korean. Translating such empty categories is also problematic other than dialogues [11], but it becomes very severe in natural conversations. A past shared task in IWSLT [12] included translator-assisted dialogues in a travel domain. A Chinese-English-Japanese corpus related to Olympic games, a.k.a. HIT corpus [13], which were also used for IWSLT shared task [14], also included some dialogues in a travel domain. These travel domain corpora have been widely used for spoken language translation studies, but these dialogues are in very limited situations and not necessarily natural conversations.

We focus on different types of dialogues called attentive listening, where a listener listens to people attentively about what they think. Conversations in attentive listening are not task-oriented so it is not easy to assume pre-defined information that can help to understand and translate them.

Table 5: Corpus statistics in the numbers of utterances (excluding backchannel and filler ones) and words. #words is based on tokenization using KyTea (ja) and Moses tokenizer (en).

	#utt.	#words (ja)	#words (en)
dev. (#1-#5)	1,476	25,780	16,235
test (#6-11)	1,510	31,857	20,099

3.2. Data

In-domain development and test data are based on the attentive listening corpus developed in NAIST [15], whose recorded and transcribed dialogues were originally in Japanese and then translated into English. We chose eleven dialogues for this task including 2,986 utterances, excluding 2,904 utterances just with backchannel and fillers. The translators were asked to translate literally with least supplement of empty categories by pronouns that were required grammatically. They could also refer to the original dialogue transcriptions with backchannel and fillers for taking the dialogue context into account.

In the recorded dialogues, many participants spoke Kansai dialect of Japanese. This caused some difficulties on Japanese morphological analyses and translation. We conducted rewriting of such expressions into standard Japanese by four annotators.

Table 5 shows the statistics of the development and test data. Since there are no other in-domain resources for this task, we did not provide any training data; participants can use any external Japanese-English resources.

3.3. Evaluation

Unfortunately we received no submissions for this task while some task registrations were made. The development and evaluation data can be obtained from the evaluation campaign website[3] for future studies.

4. Lecture Task

4.1. Definition

The lecture task covered two tracks: ASR and SLT. In the ASR track, the participants should transcribe the English and German audio. In the SLT track, these transcriptions should be translated into the other language.

[3] https://sites.google.com/site/iwsltevaluation2017/Dialogues-task

4.1.1. Data

The evaluation data for the lecture task (*tst2017*) consists of German and English recordings of talks and lectures.

The English data that participants were asked to recognize and translate consists in part of TED talks as in the years before, and in part of real-life lectures and talks that have been mainly recorded in lecture halls at KIT and Carnegie Mellon University. TED talks are challenging due to their variety in topics, but are very benign as they are very thoroughly rehearsed and planned, leading to easy to recognize and translate language. The real-life lectures that we included in the test set are more difficult to process as reflected by the scores on them in comparison to the scores on the TED talks. As this is the first edition in which we offer real-life lectures, and the amount of available test data is limited, we included both, TED talks and real lectures in the English evaluation data.

The German data consisted solely of German real-live lectures given at KIT.

4.1.2. ASR

In the ASR track participants were asked to recognize the unsegmented audio of the lectures and transcribe them automatically into the spoken word sequence. The training data for the acoustic model was limited to publicly available data, while the training data for the language model was restricted to a known list of corpora. But participants could suggest corpora to include in the list.

4.1.3. SLT

The SLT track covered the translations of university lectures and TED talks from English to German and the translation of university lectures from German to English. The participants should translate from the English and German audio signal. The challenge of this translation task is the necessity to deal with automatic, and in general error prone, transcriptions of the audio signal, instead of correct human transcriptions. Furthermore, for the lecture tasks no manual segmentation into sentences was provided. Therefore, participants needed to develop methods to automatically segment the automatic transcript and insert punctuation marks.

4.2. Evaluation

Participants to the ASR evaluation had to submit the results of the recognition of the *tst2017* sets in CTM format. The word error rate was measured case-insensitive. After the end of the evaluation, scoring was performed with the references derived from the subtitles of the TED talks and human transcripts of the real lectures.

For the SLT evaluation, participants could choose to either use their own ASR technology, or to use ASR output provided by the conference organizers.

For both input languages, the ASR output provided by the organizers was a single system output from one of the submissions to the ASR track.

Since the participants needed to segment the input into sentences, the segmentation of the reference and the automatic translation was different. In order to calculate the automatic evaluation metric, we needed to realign the sentences of the reference and the automatic translation. This was done by minimizing the WER between the automatic translation and reference as described in [16].

4.3. Submissions

We received two primary submissions for every SLT task and one primary submission for the ASR task.

4.4. Results

The detailed results of the automatic evaluation in terms of BLEU and WER can be found in Appendix B.

5. Human Evaluation

This year human evaluation focused on Multilingual translation (see Section 2) and was specifically carried out on the four language directions for which also the Zero-Shot translation task was proposed, *i.e.* $NlDe$, $DeNl$, $RoIt$ and $ItRo$.

For these four tasks, we received multilingual submissions for all the training data conditions offered, namely *large data* (ML LD), *small data* (ML SD), and *zero-shot* (ML ZS). Since multilingual translation was offered for the first time as an IWSLT task, we were interested in comparing the results with the traditional bilingual (BL) approach, where a different system is created for each language direction. For this reason, for the $NlDe$ and $RoIt$ tasks we asked those teams who participated with both ML SD and ML ZS runs to provide additional BL SD runs, to be manually evaluated as well.

A major novelty with respect to previous campaigns is that human evaluation was extended to include two different assessment methodologies, namely *direct assessment* (DA) of absolute translation quality as well as the traditional IWSLT evaluation based on *post-editing*

(PE), where the MT outupts are post-edited (*i.e.* manually corrected) by professional translators and then evaluated according to TER-based metrics [17].

We believe that carrying out a double evaluation on the same data adds great value to IWSLT 2017, since it allows to compare complementary methodologies which address different human perspectives. Indeed, while DA focuses on the generic assessment of overall translation quality, PE-based evaluation reflects a real application scenario – the integration of MT in Computer-Assisted Translation (CAT) tools – and directly measures the utility of a given MT output to translators. Also, this evaluation is particularly suitable for performing fine-grained analyses, since it produces a set of edits pointing to specific translation errors.

In this year's campaign, all systems submitted to the $NlDe$, $DeNl$, $RoIt$ and $ItRo$ tasks were officially evaluated and ranked according to DA, while PE-based evaluation was carried out on a subset of systems submitted to the $NlDe$ and $RoIt$ tasks, with the aim of analysing in detail the feasibility of the novel multilingual - and zero-shot - approach.

The human evaluation (HE) dataset created for each language direction was a subset of the corresponding 2017 test set *(tst2017)*. All the four *tst2017* sets ($NlDe$, $DeNl$, $RoIt$ and $ItRo$) are composed of the same 10 TED Talks, and around the first half of each talk was included in the HE set. The resulting HE sets are identical and include 603 segments, corresponding to around 10,000 words words for each source text.

In the following subsections we present the two evaluation methodologies and their outcomes on the HE datasets.

5.1. Direct Assessment

Recently, there has been increased interest in human evaluation of machine translation output using *direct assessment* (DA). Here, the annotator sees a simple annotation interface which shows 1) the reference translation, 2) a single candidate translation, and 3) a slider to score the translation quality from 1 to 100, focusing on the adequacy of the given translation output, compared to the given reference translation. For this year's IWSLT, we follow the setup of WMT17 [18] and run a human evaluation campaign based on DA.

Considering that any reference-based approach to evaluation will inevitably have problems when the reference translation has quality issues or a given candidate translation has an extremely different syntactic structure compared to the given reference (and might thus be judged as poor quality), we also focused on *source-based direct assessment*. This is more difficult to use as it requires a pool of bilingual annotators but (if those annotators are available) it allows to collect annotations on the actual semantic transfer between source and target languages.

Given that source-based DA eliminates reference bias and quality issues by design, we decided to run two separate DA campaigns for IWSLT, one based on the reference-based implementation of DA (identical to what has been used for WMT17) and one based on source-based DA. We used the Appraise framework [19] for both campaigns.

5.1.1. Data Preparation

Data was prepared based on the full set of 603 candidate translations used for the post-editing evaluation. However, as we wanted to ensure that each task is annotated by two annotators, we opted to randomly sample half of the candidate translations for the DA campaigns. Both source-based and reference-based direct assessment data has been prepared using the same random seed so that the only difference between the resulting tasks is in the type of "visual reference" shown to the annotator. Display order of segments and systems is identical across the campaign types.

5.1.2. Annotation Campaign

We collected annotations from a=22 annotators for $NlDe$ and $RoIt$. These language pairs contained a total of n=12 different systems and we conducted the evaluation on t=55 tasks with r=2 redundancy, so that each annotator ended up completing a total of five tasks. For $DeNl$ and $ItRo$ there were a total of a=16 annotators for

Table 6: $NlDe$ Source-based DA Human evaluation results showing average raw DA scores (Ave %) and average standardized scores (Ave z), lines between systems indicate clusters according to Wilcoxon rank-sum test at p-level $p \leq 0.05$.

#	Ave %	Ave z	**System**	**Cond.**
1	70.2	0.173	KIT	ML LD
2	70.2	0.145	KYOTO	BL SD
	69.4	0.139	KYOTO	ML SD
3	68.1	0.110	KIT	ML SD
4	68.4	0.103	KYOTO	ML ZS
	66.5	0.040	GTCT	ML ZS
	67.0	0.029	UDSDFKI	ML SD
5	64.5	-0.045	FBK	BL SD
	63.5	-0.078	UDSDFKI	ML ZS
	63.3	-0.079	FBK	ML SD
6	60.0	-0.212	FBK	ML ZS
7	57.2	-0.338	UDSDFKI	BL SD

Table 7: $NlDe$ Reference-based DA Human evaluation results showing average raw DA scores (Ave %) and average standardized scores (Ave z), lines between systems indicate clusters according to Wilcoxon rank-sum test at p-level $p \leq 0.05$.

#	Ave %	Ave z	**System**	**Cond.**
1	64.2	0.121	KIT	ML LD
2	63.5	0.100	KYOTO	ML SD
3	64.6	0.102	KYOTO	BL SD
4	63.0	0.069	KYOTO	ML ZS
	62.1	0.061	KIT	ML SD
	62.7	0.045	UDSDFKI	ML SD
	61.2	0.014	GTCT	ML ZS
5	61.1	0.017	FBK	BL SD
6	59.2	-0.076	UDSDFKI	ML ZS
	58.0	-0.092	FBK	ML SD
7	56.2	-0.178	FBK	ML ZS
	54.9	-0.241	UDSDFKI	BL SD

Table 8: $RoIt$ Source-based DA Human evaluation results showing average raw DA scores (Ave %) and average standardized scores (Ave z), lines between systems indicate clusters according to Wilcoxon rank-sum test at p-level $p \leq 0.05$.

#	Ave %	Ave z	**System**	**Cond.**
1	74.8	0.222	KYOTO	BL SD
2	74.4	0.200	KIT	ML SD
	72.1	0.131	KYOTO	ML SD
3	72.1	0.136	KYOTO	ML ZS
	71.8	0.115	KIT	ML LD
4	71.1	0.081	UDSDFKI	ML SD
	70.3	0.049	FBK	ML SD
	69.1	0.017	GTCT	ML ZS
	68.5	0.000	FBK	BL SD
5	66.9	-0.090	UDSDFKI	ML ZS
6	61.6	-0.268	FBK	ML ZS
7	55.3	-0.546	UDSDFKI	BL SD

n=9 individual systems. We annotated a set of t=40 tasks, again using r=2 redundancy, for the same annotator work load of five tasks. Our annotators were experienced linguistic consultants.

5.1.3. Results

Table 6 includes source-based DA results for $NlDe$ and Table 7 shows corresponding results from the reference-based DA campaign. Clusters are identified by grouping systems together according to which systems significantly outperform all others in lower ranking clusters, according to Wilcoxon rank-sum test. Tables 8 and 9 show results for source-based and reference-based DA for $RoIt$, respectively. Results for $DeNl$ and $ItRo$ are given in Tables 10, 11, 12, and 13.

Table 9: $RoIt$ Reference-based DA Human evaluation results showing average raw DA scores (Ave %) and average standardized scores (Ave z), lines between systems indicate clusters according to Wilcoxon rank-sum test at p-level $p \leq 0.05$.

#	Ave %	Ave z	**System**	**Cond.**
1	59.9	0.169	KIT	ML SD
2	59.9	0.162	KYOTO	ML SD
3	58.9	0.126	KYOTO	BL SD
	58.6	0.126	KYOTO	ML ZS
	58.3	0.102	KIT	ML LD
4	58.3	0.086	UDSDFKI	ML SD
5	55.2	0.014	GTCT	ML ZS
	55.1	-0.010	FBK	ML SD
	54.0	-0.045	FBK	BL SD
	54.0	-0.047	UDSDFKI	ML ZS
6	49.0	-0.190	FBK	ML ZS
7	42.9	-0.423	UDSDFKI	BL SD

Table 10: $DeNl$ Source-based DA Human evaluation results showing average raw DA scores (Ave %) and average standardized scores (Ave z), lines between systems indicate clusters according to Wilcoxon rank-sum test at p-level $p \leq 0.05$.

#	Ave %	Ave z	**System**	**Cond.**
1	70.3	0.128	KYOTO	ML ZS
2	70.0	0.088	KIT	ML LD
3	69.8	0.094	KYOTO	ML SD
	67.5	0.015	GTCT	ML ZS
	67.5	-0.002	KIT	ML SD
	67.4	-0.006	FBK	ML SD
4	66.5	-0.022	UDSDFKI	ML SD
	66.0	-0.073	UDSDFKI	ML ZS
5	62.4	-0.180	FBK	ML ZS

Note how reference-based DA scores are generally lower than those for source-based DA. It seems that given a reference, annotators are more likely to penalize a candidate translation for missing data. For the source-based case, they seem to be more focused on the actual transfer from source into target language. More detailed investigation is required to draw conclusions here and will be left for future work.

Generally, source-based and reference-based DA produce similar clusters. The decision which direct assessment to use hence comes down to the availability of bilingual annotators. If available, it seems preferable to opt for source-based DA.

For $NlDe$, KIT (ML LD) wins for both source-based and reference-based DA, with KYOTO (BL SD and ML SD) reaching second and third place. KIT is significantly better than all other systems for this language pair. Both DA methods agree on the ranking of the lower scor-

Table 11: $DeNl$ Reference-based DA Human evaluation results showing average raw DA scores (Ave %) and average standardized scores (Ave z), lines between systems indicate clusters according to Wilcoxon rank-sum test at p-level $p \leq 0.05$.

#	Ave %	Ave z	**System**	**Cond.**
1	57.7	0.126	KIT	ML LD
2	57.7	0.119	KYOTO	ML SD
	56.6	0.090	KYOTO	ML ZS
3	54.7	0.004	KIT	ML SD
4	54.4	0.009	GTCT	ML ZS
	53.7	-0.022	UDSDFKI	ML SD
	53.4	-0.068	UDSDFKI	ML ZS
	52.6	-0.073	FBK	ML SD
5	50.2	-0.156	FBK	ML ZS

Table 12: $ItRo$ Source-based DA Human evaluation results showing average raw DA scores (Ave %) and average standardized scores (Ave z), lines between systems indicate clusters according to Wilcoxon rank-sum test at p-level $p \leq 0.05$.

#	Ave %	Ave z	**System**	**Cond.**
1	77.3	0.214	KIT	ML LD
	76.5	0.189	KYOTO	ML SD
	75.9	0.173	KIT	ML SD
	74.7	0.136	KYOTO	ML ZS
2	72.6	0.048	UDSDFKI	ML SD
3	69.6	-0.070	FBK	ML SD
4	68.5	-0.103	UDSDFKI	ML ZS
	68.1	-0.115	GTCT	ML ZS
5	60.4	-0.385	FBK	ML ZS

ing systems.

For $RoIt$, KYOTO (BL SD) wins for source-based DA while KIT (ML SD) performs best for the reference-based DA campaign. For reference-based eval, the KYOTO systems drops to the third cluster. As average scores are really close across the reference-based systems, this should be investigated more. Again, both DA methods agree on the worst clusters.

For $DeNl$, we see the ML ZS system from KYOTO win over an ML LD system from KIT. While this does not happen for the reference-based campaign, the ML ZS system achieves second place there. This indicates that ML ZS can be competitive and outperforms the other approaches.

Finally, for $ItRo$ we observe identical clusters for both DA methods. Of course, average % scores and z scores differ, but the respective pairwise comparisons end up the same. Four systems achieve first rank: KIT (ML SD and ML LD) as well as KYOTO (ML SD and ML ZS).

Table 13: $ItRo$ Reference-based DA Human evaluation results showing average raw DA scores (Ave %) and average standardized scores (Ave z), lines between systems indicate clusters according to Wilcoxon rank-sum test at p-level $p \leq 0.05$.

#	Ave %	Ave z	**System**	**Cond.**
1	66.1	0.165	KIT	ML SD
	65.4	0.145	KYOTO	ML ZS
	65.1	0.142	KIT	ML LD
	64.2	0.112	KYOTO	ML SD
2	61.5	0.021	UDSDFKI	ML SD
3	60.0	-0.050	FBK	ML SD
4	58.1	-0.095	UDSDFKI	ML ZS
	58.3	-0.102	GTCT	ML ZS
5	54.0	-0.229	FBK	ML ZS

5.2. Post-Editing

5.2.1. Evaluation Data

This year, human evaluation based on post-editing was carried out on two language directions, namely $NlDe$ and $RoIt$.

In order to analyze at best the multilingual approach and to properly compare the different data conditions tested in the campaign, we selected for post-editing the six runs of the three teams who submitted both ML SD and ML ZS systems (*i.e.* KYOTO, FBK, UDSDFKI). In addition, we included in the evaluation their three unofficial BL SD runs that they were requested to submit for comparison purposes.

For each language direction, the output of the selected 9 systems on the HE set was assigned to professional translators to be post-edited (for all the details about data preparation and post-editing see [20, 21, 22]).

The resulting evaluation data consists of nine new reference translations for each of the sentences in the HE set. Each one of these references represents the *targeted translation* of the system output from which it was derived, while the post-edits of the other 8 systems are available for evaluation as additional references.

5.2.2. Results

The outcomes for the two language directions are presented in Tables 14 and 15, where systems are grouped by data condition (ML ZS, ML SD, ML LD, and BL SD). Results are analyzed according to multi-reference TER (mTER), where TER is computed against all the 9 available post-edits. Previous IWSLT PE-based evaluations demonstrated that mTER allows a more reliable and consistent evaluation of the real overall MT system performance with respect to HTER – where

TER is calculated against the targeted reference only.

Furthermore, figures are given for HTER as well as TER – both on the HE set and on the full test set – calculated against the official reference translation used for automatic evaluation (see Section 2 and Appendix A).[4] In the tables, BL SD runs are highlighted in light gray to distinguish them from the official IWSLT runs. Also, results for those official IWSLT runs that were not post-edited are given for completeness (*i.e.* KIT, GTCT). Those runs are highlighted in dark gray to signal that they are not directly comparable with the other runs: although they are evaluated with mTER on all nine available references, they do not have their corresponding targeted reference, which could result in a penalizing score.

Finally, the statistical significance of the observed differences between the systems was assessed with the *approximate randomization* method [23], a statistical test well-established in the NLP community [24] and that, especially for the purpose of MT evaluation, has been shown [25] to be less prone to type-I errors than the bootstrap method [26]. In this study, the approximate randomization test was based on 10,000 iterations. Tables 14 and 15 present the results of the test focusing on the systems within the same data condition. Information about the significance of the differences between the systems developed by the same team are given in the following discussion of results.

Table 14: $NlDe$ TED Talk task (HE *tst2017*): human evaluation results. Scores are given in percentage (%). The number next to the mTER score identifies the system(s) within the same setup w.r.t. which the difference is statistically significant at $p < 0.01$.

Cond.	**System**	**mTER HE Set 9 PErefs**	**HTER** HE Set tgt PEref	**TER** HE Set ref	**TER** Test Set ref
ML ZS	GTCT	25.36	–	64.40	65.17
	KYOTO[1]	20.33$^{(2,3)}$	25.72	64.33	64.33
	FBK[2]	26.19	33.13	67.01	67.05
	UDSDFKI[3]	27.36	33.60	68.65	68.36
ML SD	KYOTO	20.38$^{(3)}$	25.05	62.99	63.39
	FBK	21.68	27.68	65.48	65.25
	UDSDFKI	23.94	30.75	66.76	66.34
	KIT	21.34	–	62.12	62.56
ML LD	KIT	19.03	–	61.08	61.33
BL SD	KYOTO	20.31$^{(2,3)}$	26.26	63.61	63.81
	FBK	23.71$^{(3)}$	30.18	65.34	66.09
	UDSDFKI	30.27	37.25	70.72	70.30

[4]Note that since TER is an edit-distance measure, lower numbers indicate better performance.

Table 15: $RoIt$ TED Talk task (HE *tst2017*): human evaluation results. Scores are given in percentage (%). The number next to the mTER score identifies the system(s) within the same setup w.r.t. which the difference is statistically significant at $p < 0.01$.

Cond.	**System**	**mTER HE Set 9 PErefs**	**HTER** HE Set tgt PEref	**TER** HE Set ref	**TER** Test Set ref
ML ZS	GTCT	26.94	–	61.80	61.11
	KYOTO[1]	22.65$^{(2,3)}$	29.33	60.58	60.26
	FBK[2]	29.16	37.38	64.21	63.32
	UDSDFKI[3]	28.74	35.79	64.79	63.97
ML SD	KYOTO	20.27	27.17	60.14	59.75
	FBK	20.74	29.01	60.45	59.65
	UDSDFKI	23.39	31.25	61.95	60.77
	KIT	22.81	–	58.70	58.29
ML LD	KIT	22.48	–	58.46	57.87
BL SD	KYOTO	18.39$^{(2,3)}$	26.09	58.90	58.55
	FBK	22.69$^{(3)}$	30.34	61.25	60.73
	UDSDFKI	26.73	34.85	61.74	63.40

Looking at the tables, some conclusions can be drawn about the feasibility of multilingual MT. It is interesting to note that the same considerations hold across language directions – although to varying degrees. First of all, the impressive results of ML SD runs show that multilingual systems are indeed an effective alternative to traditional bilingual systems. Even more noticeably, ML ZS systems are able to reach a reasonably good quality also when faced with such an extreme translation scenario, clearly showing the feasibility of the zero-shot approach. Finally, by comparing the systems' performance within each condition, some specific characteristics of the ML and BL approach emerge. As we can see in the tables, the three BL SD systems are all significantly different, while ML SD systems (and ML ZS, although to a lesser extent) are mostly similar to each other.

We now compare in detail the systems produced by each team in the different conditions. Considering the $NlDe$ direction (Table 14), KYOTO provides the clearest demonstration of the feasibility of the multilingual zero-shot approach, since it obtains the same outstanding results in all the three translation conditions. FBK and UDSDFKI systems show a very similar behaviour. They further confirm the effectiveness of the multilingual approach, since their ML SD runs improve over their corresponding BL SD runs, and with a statistically significant difference. As for zero-shot translation, FBK and UDSDFKI systems still show a reasonably good quality, although results are significantly lower

than those obtained in the ML SD data condition (+4.51 mTER points for FBK and +3.42 for UDSDFKI). With respect to the BL SD runs, UDSDFKI ML ZS performance is higher (though the difference is not statistically significative), while FBK ML ZS results are significantly lower.

Regarding non-comparable runs (in dark grey in the table), we see that the ML ZS system developed by GTCT is in line with the other results. As for KIT, its performance on the ML LD data condition confirms that using more data for training can help improving results. However, the difference with respect to its corresponding ML SD system is not particularly remarkable, although statistically significant.

It is worthwhile to note that the differences between systems highlighted by mTER scores are not so marked when looking at TER scores. As also shown in previous IWSLT evaluations, TER calculated against one independent reference does not allow to discriminate properly between systems; this study supports once more the need for human evaluation to shed light on the peculiarities of the systems.

Considering the $RoIt$ language direction (Table 15), we can draw the same conclusions about the feasibility of the multilingual approach, although results for the zero-shot task are less notable. KYOTO ML SD system is not significantly different from the traditional BL SD system, even though it does not reach its performance. On the contrary, results for ML ZS system are significantly lower than those obtained by the ML SD one, although the difference is only 2.38 mTER points.

As seen for the $NlDe$ direction, FBK and UDSDFKI ML SD runs significantly improve over their corresponding BL SD runs; however, for the $RoIt$ direction the drop in performance of the ML ZS systems with respect to the ML SD ones is more critical (8.42 mTER points for FBK and 5.35 for UDSDFKI). Also, ML ZS runs are worse than BL SD runs, even though for UDSDFKI the difference is not statistically significant.

5.3. Future Work

We intend to run a deeper analysis on the human evaluation corpus created as part of IWSLT. Not only does it make sense to more closely investigate the differences of source-based and reference-based DA, but it will also be very interesting to compare the results of such "general quality focused" annotation work to more targeted approaches such as post-editing. As we do have such data for two of the language pairs, the resulting three-way dataset will be released for future research.

6. Conclusions

This year the IWSLT Evaluation Campaign featured three tasks: the Multilingual task, evaluating single MT systems translating across multiple languages, the Dialogues task, addressing MT of human-to-human dialogues, and the Lecture task, targeting speech transcription and translation of real-life university lectures. This paper overviews the structure of each task, its experimental conditions, the training and evaluation data made available, and reports on its participation and main outcomes. Besides documenting the evaluation campaign to the perusal of the workshop participants, we hope that this paper will also be useful to researchers and practitioners interested in using our evaluation benchmarks in the future.

7. Acknowledgements

Human evaluation based on post-editing and part of the work by FBK's authors were supported by the CRACKER project, which receives funding from the EU's Horizon 2020 research and innovation programme under grant agreement no. 645357.

8. References

[1] N. Ruiz and M. Federico, "Complexity of spoken versus written language for machine translation," in *Proc. of EAMT*, Dubrovnik, Croatia, 2014, pp. 173–180.

[2] S. M. Lakew, Q. F. Lotito, M. Turchi, M. Negri, and M. Federico, "FBK's multilingual neural machine translation system for IWSLT 2017," in *Proc. of IWSLT*, Tokyo, Japan, 2017.

[3] C. Bei and H. Zong, "Towards better translation performance on spoken language," in *Proc. of IWSLT*, Tokyo, Japan, 2017.

[4] N.-Q. Pham, M. Sperber, E. Salesky, T.-L. Ha, J. Niehues, and A. Waibel, "KIT's multilingual neural machine translation systems for IWSLT 2017," in *Proc. of IWSLT*, Tokyo, Japan, 2017.

[5] R. Dabre, F. Cromieres, and S. Kurohashi, "Kyoto university MT system description for IWSLT 2017," in *Proc. of IWSLT*, Tokyo, Japan, 2017.

[6] P. Bahar, J. Rosendahl, N. Rossenbach, and H. Ney, "The RWTH Aachen machine

translation systems for IWSLT 2017," in *Proc. of IWSLT*, Tokyo, Japan, 2017.

[7] P. Przybysz, M. Chochowski, R. Sennrich, B. Haddow, and A. Birch, "The Samsung and University of Edinburgh's submission to IWSLT17," in *Proc. of IWSLT*, Tokyo, Japan, 2017.

[8] C. España-Bonet and J. van Genabith, "Going beyond zero-shot MT: combining phonological, morphological and semantic factors. The UdS-DFKI system at IWSLT 2017," in *Proc. of IWSLT*, Tokyo, Japan, 2017.

[9] M. Cettolo, C. Girardi, and M. Federico, "WIT^3: Web inventory of transcribed and translated talks," in *Proc. of EAMT*, Trento, Italy, 2012.

[10] S. Takeno, M. Nagata, and K. Yamamoto, "Integrating empty category detection into preordering machine translation," in *Proc. WAT*, Osaka, Japan, 2016.

[11] T. Chung and D. Gildea, "Effects of empty categories on machine translation," in *Proc. of EMNLP*, Cambridge, US-MA, 2010.

[12] M. Paul, "Overview of the IWSLT 2009 evaluation campaign," in *Proc. of IWSLT*, Tokyo, Japan, 2009.

[13] M. Yang, H. Jiang, T. Zhao, and S. Li, "Construct trilingual parallel corpus on demand," in *Chinese Spoken Language Processing. Lecture Notes in Computer Science*, Q. Huo, B. Ma, E.-S. Chng, and H. Li, Eds. Springer, Berlin, Heidelberg, 2006, vol. 4274, pp. 760–767.

[14] M. Federico, M. Cettolo, L. Bentivogli, M. Paul, and S. Stüker, "Overview of the IWSLT 2012 evaluation campaign," in *Proc. of IWSLT*, Hong Kong, 2012.

[15] H. Tanaka, K. Yoshino, K. Sugiyama, S. Nakamura, and M. Kondo, "Multimodal interaction data between clinical psychologists and students for attentive listening modeling," in *Proc. of O-COCOSDA*, Bali, Indonesia, 2016.

[16] E. Matusov, G. Leusch, O. Bender, and H. Ney, "Evaluating machine translation output with automatic sentence segmentation," in *Proc. of IWSLT*, Pittsburgh, US-PA, 2005.

[17] M. Snover, B. Dorr, R. Schwartz, L. Micciulla, and J. Makhoul, "A study of translation edit rate with targeted human annotation," in *Proc. of AMTA*, Cambridge, US-MA, 2006.

[18] O. Bojar, R. Chatterjee, C. Federmann, Y. Graham, B. Haddow, S. Huang, M. Huck, P. Koehn, Q. Liu, V. Logacheva, C. Monz, M. Negri, M. Post, R. Rubino, L. Specia, and M. Turchi, "Findings of the 2017 conference on machine translation (WMT17)," in *Proc. of WMT: Shared Task Papers*, Copenhagen, Denmark, 2017.

[19] C. Federmann, "Appraise: An open-source toolkit for manual evaluation of machine translation output," *The Prague Bulletin of Mathematical Linguistics*, vol. 98, pp. 25–35, September 2012.

[20] M. Cettolo, J. Niehues, S. Stüker, L. Bentivogli, and M. Federico, "Report on the 11th IWSLT evaluation campaign, IWSLT 2014," in *Proc. of IWSLT*, Lake Tahoe, US-CA, 2014.

[21] M. Cettolo, J. Niehues, S. Stüker, L. Bentivogli, R. Cattoni, and M. Federico, "The IWSLT 2015 evaluation campaign," in *Proc. of IWSLT*, Da Nang, Vietnam, 2015.

[22] ——, "The IWSLT 2016 evaluation campaign," in *Proc. of IWSLT*, Seattle, US-WA, 2016.

[23] E. W. Noreen, *Computer Intensive Methods for Testing Hypotheses: An Introduction*. Wiley Interscience, 1989.

[24] N. Chinchor, L. Hirschman, and D. D. Lewis, "Evaluating message understanding systems: An analysis of the third message understanding conference (muc-3)," *Computational Linguistics*, vol. 19, no. 3, pp. 409–449, 1993.

[25] S. Riezler and J. T. Maxwell, "On some pitfalls in automatic evaluation and significance testing for MT," in *Proc. of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, Ann Arbor, US-MI, 2005.

[26] B. Efron and R. J. Tibshirani, *An Introduction to the Bootstrap*. Chapman and Hall, 1993.

Appendix A. Automatic Evaluation for the Multilingual Task

· Table scores refer to the official testset (*tst2017.mltlng*)
· *BLEU* and *TER* scores are given as percent figures (%)
· ML, BL, SD, LD and ZS stand for multilingual, bilingual, small-data, large-data and zero-shot conditions, respectively
· BL SD systems were developed by three participants on explicit request of the organizers for comparison purposes

system	cond.	BLEU	NIST	TER	BLEU	NIST	TER
		Dutch-German			**German-Dutch**		
FBK	ML SD	18.59	5.177	65.24	19.16	5.583	61.45
	ML ZS	16.96	4.931	67.04	17.17	5.297	63.25
	BL SD	17.93	5.139	66.09	–	–	–
GTCT	ML ZS	19.00	5.208	65.17	19.59	5.565	61.27
KIT	ML SD	20.47	5.542	62.56	19.77	5.735	59.37
	ML LD	21.06	5.657	61.33	20.00	5.763	59.21
KYOTO	ML SD	20.27	5.487	63.39	19.64	5.733	60.24
	ML ZS	19.68	5.368	64.33	20.31	5.751	59.99
	BL SD	19.50	5.390	63.81	19.86	5.754	59.93
UDSDFKI	ML SD	18.28	5.133	66.34	18.96	5.492	63.50
	ML ZS	16.28	4.874	68.36	17.38	5.375	62.72
	BL SD	16.43	4.767	70.30	–	–	–
		Dutch-Italian			**Italian-Dutch**		
FBK	ML SD	19.33	5.471	62.88	20.27	5.568	61.78
	ML ZS	19.76	5.422	62.99	20.00	5.548	61.91
GTCT	ML ZS	21.21	5.722	60.84	21.80	5.784	60.09
KIT	ML SD	20.41	5.599	61.64	22.14	6.005	58.34
	ML LD	20.94	5.706	60.18	21.95	6.003	58.21
KYOTO	ML SD	19.86	5.530	62.07	22.32	5.922	59.16
	ML ZS	20.74	5.602	61.85	22.76	5.911	59.16
UDSDFKI	ML SD	19.12	5.419	63.69	20.08	5.560	62.02
	ML ZS	19.39	5.435	63.68	19.88	5.563	61.92
		English-Dutch			**Dutch-English**		
FBK	ML SD	26.72	6.536	53.45	29.79	7.078	50.27
	ML ZS	26.11	6.501	54.34	30.04	7.081	50.04
GTCT	ML ZS	29.08	6.805	51.47	32.78	7.422	47.35
KIT	ML SD	29.15	6.903	51.08	31.79	7.340	47.84
	ML LD	30.22	6.984	50.45	31.95	7.399	46.88
KYOTO	ML SD	28.80	6.824	52.16	30.49	7.131	49.04
	ML ZS	30.18	6.963	50.71	30.63	7.158	48.94
UDSDFKI	ML SD	26.49	6.529	53.72	29.53	7.112	49.64
	ML ZS	26.37	6.534	54.19	29.69	7.073	50.03
		English-Italian			**Italian-English**		
FBK	ML SD	29.60	6.821	50.74	34.24	7.618	44.45
	ML ZS	28.86	6.687	51.80	34.16	7.638	44.38
GTCT	ML ZS	32.84	7.222	47.63	37.84	8.100	41.06
KIT	ML SD	32.04	7.147	48.36	36.30	7.945	41.97
	ML LD	32.32	7.219	48.11	36.46	7.980	41.89
KYOTO	ML SD	30.79	6.921	50.48	34.73	7.631	45.07
	ML ZS	30.99	6.989	49.69	35.28	7.679	44.51
UDSDFKI	ML SD	29.62	6.855	50.48	33.77	7.644	44.07
	ML ZS	29.68	6.849	50.55	33.77	7.596	44.71
		German-Italian			**Italian-German**		
FBK	ML SD	16.84	5.094	65.67	16.88	4.92	68.38
	ML ZS	16.28	4.971	66.76	16.13	4.828	69.22
GTCT	ML ZS	18.56	5.363	63.44	18.09	5.091	67.28
KIT	ML SD	17.79	5.265	63.81	19.32	5.344	64.71
	ML LD	18.04	5.280	63.01	19.85	5.414	64.16
KYOTO	ML SD	17.54	5.262	64.32	19.10	5.339	64.73
	ML ZS	17.67	5.227	64.77	19.20	5.287	65.31
UDSDFKI	ML SD	16.66	5.096	66.12	16.48	4.870	69.15
	ML ZS	16.73	5.106	66.09	16.27	4.873	68.79

system	cond.	BLEU	NIST	TER	BLEU	NIST	TER
		Italian-Romanian			**Romanian-Italian**		
FBK	ML SD	19.06	5.155	64.87	21.34	5.811	59.65
	ML ZS	16.58	4.783	67.53	18.32	5.296	63.32
	BL SD	–	–	–	21.71	5.776	60.73
GTCT	ML ZS	18.62	5.027	65.54	20.39	5.573	61.11
KIT	ML SD	21.08	5.566	61.31	22.54	6.0209	58.28
	ML LD	21.09	5.629	60.68	22.35	6.013	57.87
KYOTO	ML SD	20.60	5.446	62.76	21.89	5.820	59.75
	ML ZS	20.37	5.385	62.79	21.85	5.789	60.26
	BL SD	–	–	–	23.14	6.026	58.55
UDSDFKI	ML SD	17.77	5.001	66.40	21.22	5.743	60.77
	ML ZS	16.07	4.752	68.21	18.67	5.352	63.97
	BL SD	–	–	–	18.94	5.345	63.40
		Dutch-Romanian			**Romanian-Dutch**		
FBK	ML SD	16.54	4.759	68.32	18.92	5.396	63.48
	ML ZS	15.88	4.698	68.57	17.72	5.272	64.51
GTCT	ML ZS	18.11	4.966	66.55	20.02	5.586	61.87
KIT	ML SD	17.43	5.067	64.98	19.28	5.674	60.93
	ML LD	17.52	5.103	64.48	19.19	5.645	61.10
KYOTO	ML SD	17.65	5.055	65.84	20.24	5.745	60.90
	ML ZS	17.74	5.056	65.75	20.47	5.699	61.14
UDSDFKI	ML SD	14.83	4.529	71.33	17.58	5.281	65.16
	ML ZS	14.93	4.532	71.79	17.26	5.286	64.44
		English-German			**German-English**		
FBK	ML SD	20.88	5.501	63.50	25.62	6.528	54.05
	ML ZS	20.67	5.471	63.80	25.22	6.453	54.54
GTCT	ML ZS	23.08	5.861	60.63	28.04	6.851	51.42
KIT	ML SD	23.86	6.029	59.22	26.76	6.694	52.43
	ML LD	25.49	6.212	57.75	27.47	6.803	51.26
KYOTO	ML SD	23.25	5.924	60.23	26.45	6.609	52.65
	ML ZS	23.63	5.936	60.22	27.08	6.678	52.49
UDSDFKI	ML SD	20.63	5.535	63.37	24.75	6.445	54.74
	ML ZS	20.20	5.504	63.49	24.54	6.442	55.22
		English-Romanian			**Romanian-English**		
FBK	ML SD	21.95	5.600	61.40	28.93	6.964	49.91
	ML ZS	21.54	5.575	61.41	28.52	6.925	50.57
GTCT	ML ZS	23.89	5.906	58.81	31.79	7.368	47.22
KIT	ML SD	25.09	6.132	56.92	30.71	7.208	48.18
	ML LD	25.25	6.133	56.95	30.69	7.242	48.01
KYOTO	ML SD	24.66	6.059	57.70	29.58	7.063	49.10
	ML ZS	24.49	6.073	57.16	30.23	7.102	48.78
UDSDFKI	ML SD	20.35	5.425	63.30	27.99	6.877	51.44
	ML ZS	20.25	5.353	63.99	28.25	6.902	51.09
		German-Romanian			**Romanian-German**		
FBK	ML SD	14.62	4.479	70.96	15.87	4.762	69.04
	ML ZS	13.93	4.400	71.10	15.47	4.695	69.87
GTCT	ML ZS	16.23	4.689	69.04	17.95	5.057	67.03
KIT	ML SD	14.99	4.690	67.59	18.01	5.181	66.01
	ML LD	15.31	4.737	67.12	18.14	5.198	65.44
KYOTO	ML SD	16.27	4.794	68.08	17.94	5.135	66.44
	ML ZS	16.08	4.822	67.76	18.40	5.152	66.24
UDSDFKI	ML SD	13.89	4.381	72.13	15.30	4.667	71.66
	ML ZS	13.83	4.287	72.97	15.01	4.652	71.37

Appendix B. Automatic Evaluation for the Lecture Task

ASR: Talk English and German
Results in Word Error Rate (WER)

German		English	
Testset	KIT	**Testset**	KIT
lecture 01	16.6	lecture 01	9.9
lecture 03	31.8	lecture 02	11.7
lecture 04	17.7	ted 2403	6.6
		ted 2429	10.6
		ted 2438	6.6
		ted 2439	15.5
		ted 2440	4.1
		ted 2442	6.7
		ted 2447	6.0
		ted 2507	6.2
All lectures	21.3	All lectures	10.3
All ted	–	All ted	7.7
All	22.8	All	8.5

SLT: Lecture translation task
Results in BLEU

German - English			English - German		
Testset	KIT	UEDIN	**Testset**	KIT	UEDIN
lecture 01	17.31	18.86	ted 2403	18.67	16.48
lecture 03	7.66	8.39	ted 2413	17.06	13.91
lecture 04	15.32	17.58	ted 2429	23.87	16.17
			ted 2438	17.14	8.05
			ted 2439	14.95	8.71
			ted 2440	13.52	13.28
			ted 2442	20.89	16.30
			ted 2447	11.59	7.73
			ted 2478	17.67	12.69
			ted 2507	16.64	14.15
			lecture 01	23.40	23.56
			lecture 02	18.75	22.70
All	12.50	13.99	ALL	18.59	15.98