Final Evaluation Report

EU-BRIDGE Bridges Across the Language Divide

Grant number 287658 | ICT-2011.4.2 Language Technologies

Authors

Philipp Koehn, Yuqi Zhang, Christian Dugast, Juliet Gauthier, Simon Grimsey, Sarah Fuenfer, Markus Mueller, Sebastian Stüker, Volker Steinbiss

Scientific Coordinator

Alexander Waibel

The work leading to these results has received funding from the European Union under grant agreement n°287658.

02.03.2015, Karlsruhe, Germany

Collaborative Project

EU BRIDGE

Bridges Across the Language Divide

Grant number 287658 — ICT-2011.4.2 Language Technologies

Start date 01 February 2012 / duration 36 Months

D6.3: Final Evaluation Report

Type of Activity	RTD
Contributing WP(s)	WP6
Nature	Report
Distribution	Public
Contractual date of delivery	January 31, 2015
Actual date of delivery	March 02, 2015
Date of last update	February 13, 2015
Status & version	Final
Number of pages	90
WP / Task responsible	WP 6 / UEDIN
Authors	Philipp Koehn, Yuqi Zhang, Christian Dugast, Juliet Gauthier, Simon Grimsey, Sarah Fuenfer, Markus Mueller, Sebastian Stüker, Volker Steinbiss
Reviewers	Sebastian Stüker, Roma Wiezorek
Keywords	WP6, evaluation, field test

For copies of reports, updates and project activities and other EU-BRIDGE related information, contact:

Margit Rödder
roedder@kit.edu
KIT – Press and Communication
Adenauerring 2, Building 50.20
76131 Karlsruhe, Germany
Tel.: +49 721 608 48676

Copies and public reports and other material can also be accessed via the project's homepage: http://www.eu-bridge.eu/

March 2, 2015

Part of the Seventh Framework Programme

Funded by the EC - DG CONNECT

Executive Summary

This document contains information about the work carried out in WP 6. The main objective of this work package is to evaluate the technology developed by the project and to test it in practical use. The specific objectives of the work package are: (1) internal evaluation on standardised test sets, (2) participation in external evaluation campaigns, and (3) field testing in commercial settings.

Contents

1 Task 6.1: Internal evaluation on standardised test sets

In order to ensure that the performance of the workers developed in Work Package 3, that are at the core of the use cases realised in Work Package 5, is sufficient for the use cases that they are supposed to be used in, we performed several internal evaluations. In the last year of EU-BRIDGE internal evaluations focused on the subtitling use case and the European Parliament use case.

Subtitling—Euronews and Sky News The main essential step in subtitling, and arguably its most important performance bottleneck, is the Automatic Speech Recogniser (ASR), which converts speech audio into a raw sequence of textual words. To assess the quality that can be achieved in this step on relevant real-world data, as well as drive progress towards ever lower error rates, the EU-BRIDGE project embraced the concept of 'coopetition'. This is a paradigm under which multiple ASR systems are trained and set up independently from each other to compete on a certain well-defined task. Afterwards, when the results are known, everybody has the opportunity to learn and benefit from the relative strengths and weaknesses of the participating systems.

European Parliament For the European Parliament a interpreter support tool is being developed which is based upon reliable terminology extraction workers and named entity taggers. Therefore these two technologies were evaluated in domain specific, internal evaluations.

1.1 Euronews Evaluation

Euronews is a company broadcasting news in several languages through two main channels: a satellite TV and a web portal. In 2013 the EU-BRIDGE consortium made an agreement with Euronews for the exchange of data inside the consortium. Following this agreement, a multilingual data set was prepared that allowed to build comparable corpora for AM training, ASR development and evaluation for 10 EU-BRIDGE languages: Arabic, English, French, German, Italian, Polish, Portuguese, Russian, Spanish and Turkish.

In order to prepare material for AM training, it is necessary to collect a set of audio recordings together with their orthographic transcription. Our target was to collect about 100 hours of raw speech—including silence, music, etc.—for each language. For that we used Euronews web data. To obtain a reliable transcription of each news in an automatic way, we applied a light supervised training procedure Lamel et al. (2002). The amount of data that is retained ranges from about 35% to about 60% of the material, depending on the language. The data roughly correspond to 100 hours of speech for each language but Polish, for which we collected about 60 hours of speech Gretter (2014b,a).

For testing, we collected about 2 hours from the Web and about 2 hours from the TV channel for each language. All these data were manually transcribed. In January 2014 a first dry run was organised, restricted to EU-BRIDGE partners, on a portion of that data (about half an hour for development and about half an hour for evaluation, for every language).

To train their systems, participants were allowed to use any speech and text data, respecting the following cut-off dates for AM and LM data:

- March 31, 2013 for Polish;
- June 30, 2013 for all the other languages.

In January 2015 a first evaluation was organised, with the same conditions, but using more data: about 80 minutes of speech for development and 80 minutes of speech for evaluation, for every language. The last 80 minutes will be kept for further evaluations. The following deadlines were used:

- Nov 2013: training data for AM available (100 hours per language, only Polish 60 hours)
- Dec 12, 2014: Dev set distribution (sphere + uem + trs + stm files, 80 minutes per language)
- Dec 24, 2014: Eval set distribution (sphere + uem files, 80 minutes per language)
- Jan 16, 2015: deadline for submitting results (ctm format)

Table 1 contains preliminary results of the primary submissions for all languages and partners. A baseline was provided, using only Nov 2013 training data both for AM and LM models. Linguistic normalisations were applied to different languages, for instance for Arabic the tool QCRI-normaliser.3.0 was used in order to assure diacritics normalisation.

	Dev	Eval						
	baseline	baseline	FBK	KIT	PJIIT	PEV	RWTH	UEDIN
English	23.4%	26.8%	13.3%	13.4%		21.5%		
French	23.1%	25.6%		11.2%				
Spanish	13.8%	16.0%	9.1%					
German	20.6%	20.7%	11.6%					
Italian	15.5%	15.5%	8.0%	15.5%		10.6%		
Polish	21.5%	19.1%			14.3%		6.8%	
Portuguese	35.1%	35.8%		18.0%		23.2%	19.1%	
Russian	34.1%	30.8%	16.7%	12.9%				
Arabic	37.1%	34.5%	29.3%			21.6%	21.3%	
Turkish	28.9%	31.9%	20.5%	21.7%				

Table 1: Euronews evaluation 2015. Results in terms of word error rate. Baselines were provided by FBK.

1.2 Sky News evaluation

In the first two years of the project, recognition of Weatherview data was chosen for the purpose of evaluating subtitling technologies. Weatherview data consists of three-minute talks about the weather in the UK, employing a rather limited vocabulary, spoken by a fairly limited set of speakers, and recorded under studio conditions. By the third year, however, it became clear the limits of this task were reached. Not only was its real-world applicability fairly limited, its simplicity led to exceedingly low error rates across the board, with any differences attributable to chance rather than conceptual differences between systems. Therefore, in the third year, the Weatherview task was abandoned in favour of Sky News.

This task consists of ASR on continuous recordings of news broadcasts, interspersed with advertisements. Unlike Weatherview, the speech in these tasks is occasionally overlaid with jingles or background noise. The bulk of it is spoken by a limited set of newsreaders recorded under studio conditions, but there is also a large number of interviewees, with different accents and recorded under a variety of different conditions. All these things put together make Sky News a highly challenging task for ASR. However, the potential economic benefits are large, since a high-quality transcription may increase speed and accuracy of the professional subtitlers that provide captions for it, as it is broadcast 24 hours a day, 7 days a week.

	dev set	test set
FBK	24.5	17.5
KIT	19.9	17.6
PEV	N/A	24.2
RWTH	19.1	17.5
UEDIN	22.0	20.3

Table 2: The word error rates (%) on dev and test sets for the participants in the Sky News evaluation.

1.2.1 Training data

For acoustic model training in the Sky News task, any possible data a participant deemed suitable was allowed, with the obvious exception of data included in dev and test set, see section 1.2.2. In addition, a large archive of in-domain data was made available by Red Bee Media. This archive contains close to 14,000 segments of 15 minutes each, each of which is accompanied by close caption subtitles, that can readily be used as an approximate transcription. However, this data is not integrally useful for acoustic model training. Since it includes many recordings that are made on the same day, the repetitive nature of hourly news broadcasts causes a significant amount of data duplication, that has to be dealt with in some way.

1.2.2 Test sets

The development set About three hours worth of data were taken from the archive described in Section 1.2.1, and transcribed manually. In this transcription process, the decision was made to transcribe only segments of news, and to skip the segments that contain advertising. After all, only the former is considered to be part of the task at hand. The audio files, along with segment timings, manual transcriptions, and the annotators' best guess at the identity of the speakers were provided to all participants.

The test set The test set is quite a bit smaller than the dev set, at around 50 minutes of speech. Unlike for the dev set, the test set was only provided in the form of audio. Since no segmentation, or speaker information was provided, participants had to overcome this by applying any kind of speaker diarisation at their disposal. This was meant to make the task more similar to a real-word scenario. However, as before, only segments that correspond to news are considered in the evaluation. Recognition results on segments that are labelled as advertising were disregarded in the evaluation.

1.2.3 Results

Five EU-BRIDGE partners participated in the Sky News evaluation. In alphabetical order, they are FBK (Fondazione Bruno Kessler), KIT (Karlsruhe Institute of Technology), PEV (PerVoice), RWTH (Rheinisch-Westfälische Technische Hochschule Aachen), and UEDIN (University of Edinburgh). Their resulting word error rates on both dev set and test set, are listed in Table 2. It has to be noted that when the challenge was issued, there was some ambiguity whether the use of the dev set's oracle segmentation and speaker information was allowed. As a result, some participants relied on this given information, while others considered speaker diarisation a part of the task, the same way as for the test set. It is therefore difficult to form a clear picture based on just the dev scores. For the test set, where no such confusion was present, the results are more reliable.

1.3 European Parliament

The European Parliament requires a tool to help their interpreters extract terminological expressions and highlight important numbers and dates. The tool, described in deliverable D5.4.3, has been developed and released at `https://www.interpreter-support.eu`.

The internal evaluations have been carried out on the following tasks:

- Terminology extraction
- Named-entity tagging on EP documents

1.3.1 Terminology extraction

Terminology extraction in the EU-BRIDGE project aims to extract terminological keywords and phrases from parliamentary documents such as plenary session reports, presentations and other parliamentary resources. A terminology could be a single word or a phrase with multiple words. The algorithm of extraction is described in detail in the deliverable report of Task 1.4.

The data for this task is human annotated by the EP interpreters and collected and has been split into training and test set.

Training set: The training set for terminology support are the annotated reports and preparation documents provided by interpreters. It includes 18 files and about 700 terminologies.

Test data: The test set used for terminology support are the English reports of the January 15th, 16th and 17th, 2013. Overall 13 reports have been annotated and it has 937 sentences which provides about 600 terminologies.

Evaluation The terminology extraction is evaluated with the public tool Termometer[1], which provides precision(P), recall(R), f-measure(F) and threshold precision(T-P), threshold recall(T-R) and threshold f-measure(T-F). A threshold-score counts a term correct if its string similarity with the reference term lies beyond a given threshold. The threshold used in this evaluation is 0.663, which is generated by the evaluation tool automatically based on the clustering on the reference terms.

The evaluation results are presented in Table 3 as percentages. Only the term candidates with occurrence over 1 are kept. There are 602 term references for these test sets. The basic setting is to use the tf*idf scoring method. Then C-value and NC-value are added. C-value has improved the results, especially the precisions, which is more important for the interpreters. Adding NC-value increases the precisions further, but decreases the recalls at the same time.

	T-P	T-R	T-F	P	R	F
Tf-idf	48.9	55.2	51.8	6.3	26.3	10.2
+C-value	59.2	54.1	56.5	7.6	23.9	11.6
+NC-value	63.1	43.3	51.3	11.1	18.9	14.0

Table 3: Terminology extraction results (term occurrence >1)

[1] http://sourceforge.net/projects/termometerxd/

1.3.2 Named-entity tagging

Named-Entity(NE) tagging aims to highlight important numbers, names, locations, dates etc. in sentences. KIT continued to work on improving the tagging quality on text.

Named-entity tagging on text. Overall the tagging system supports 13 named entity types which include 8 common types and 5 types that are specific to the European Parliament.

The training and test data include debates and reports from the European Parliament. The training data is manually annotated. The data is described in the Table 4. By data cleaning and increasing the training data size, the F-measure has increased for most types. The detailed results are presented in Table 5. The types Time, Rule and Resolution do not appear in the table as they do either not occur in the evaluation material or too sparsely.

	Files	Words	Terms
Train	30	296,843	11,930
Test	23	38,575	1,811

Table 4: Named-entity data

Tag	Precision[%]	Recall[%]	F-Measure[%]	F-Measure[%] last year
Number	66.7	40.4	50.3	30.2
Date	74.1	85.5	79.4	72.0
Abbreviation	58.8	87.0	70.2	57.4
Organisation	90.1	65.6	76.0	76.7
Location	77.2	82.1	79.6	91.4
Person	53.2	71.1	60.9	59.9
Money	86.7	92.9	89.7	70.0
Percent	100.0	92.9	96.3	90.0
Article	91.7	91.7	91.7	87.0
Directive	40.0	66.7	50	50

Table 5: Named-entity results

1.4 Polish-English lecture translation

An attempt was made to create an pipeline for speech translation of Polish lecture domain. This would serve not only as a benchmark, but could potentially have some real-world use at PJIIT. The typical use case entails a lecturer speaking in one language and the listeners or students following the lecture in another language. For our use, the Polish-to-English direction would make most sense and this is what was tested. The pipeline described here consists of the ASR system built using Kaldi (Povey et al., 2011) and an SMT system based around Moses (Koehn et al., 2007). We have the potential to also utilise a TTS system and create a complete speech-to-speech pipeline, but for any practical use a simple text-only output would probably be more useful, similarly to how it was extensively tested at the KIT.

1.4.1 ASR module

The corpus used to develop the ASR module consisted of various lectures collected by the Institute during this and previous projects. This included lectures from the PlatonTV[2] platform

[2] `http://tv.pionier.net.pl/`

Model	Beam	Vocab	WER
tri2b	default	69k	68.07%
FMLLR	default	69k	55.48%
tri2b	wide	69k	52.33%
FMLLR	wide	69k	41.36%
FMLLR lattice oracle	wide	69k	26.45%
tri2b	wide	214k	21.83%
FMLLR	wide	214k	**19.50%**
FMLLR lattice oracle	wide	214k	7.94%

Table 6: Results of the ASR experiments. The default beam width parameter allows real-time decoding, while the wide searches a much larger hypothesis space, giving better results at the cost of speed and memory performance.

and lectures from the Nomadic[3] project recorded at PJIIT. More data was used for developing the language models: most of the corpora used in other experiments (online sources, magazines, radio and TV), available Polish corpora (Rzeczpospolita, IPI PAN corpus), transcripts of the audio data mentioned above and a collection of undergraduate theses from the PJIIT (highly correlated with the lectures used for evaluation). More details on the LM are available in the SMT section.

Initial acoustic models were trained on over 250 hours of various audio corpora and then fine tuned on a corpus of about 25 hours of in-domain data, mentioned above. The test set consisted of about 3.5 hours of lectures. The reason for such a low number was that apart from transcribing, the data also had to be translated in order to test the complete S2S pipeline, which came at an increased cost. The dictionary of the initial language model contained about 69k words, but the amount of OOV was very high (11% of words in transcription and 27% of the dictionary), so a much larger model was trained to reduce the OOV, which came at a cost of over 214k words. This also meant that the size of the FST increased from 878 MB to over 14 GB!

The experiment was performed on two types of models in Kaldi. The first one was the baseline triphone model trained on the MFCC features transformed using LDA (aka tri2b) and the second used FMLLR to implement speaker-adapted models. Table 6 lists results for the chosen experiments performed for this task. The initial experiment used the standard beam and a smaller vocabulary which gave poor results. Increasing the beam helped considerably (at the cost of processing time), but the number of OOV left the WER still very high. Using a much larger LM improved the performance to a much more satisfying result by reducing the OOV count to a very low amount ($< 3\%$), but there were still a lot of errors remaining. An experiment was made using the lattice oracle, which looks up the path with the lowest WER within the lattices generated by the decoder. The large discrepancy between the oracle and the best-path scoring methods demonstrates the weakness of the models used to evaluate the best-path output. It is our suspicion that even though out LM contained a very low OOV rate, the rare words were so sparse in our limited data set, that they weren't modelled accurately. A solution to this, that is currently being tested, is to use a more open-vocabulary approach instead.

1.4.2 SMT module

The SMT system is based on the TED talks corpus (about 17 MB) which includes almost 2.5 million words. The transcripts in the training corpus are provided as pure text encoded with

[3] `http://nomad.pja.edu.pl/`

UTF-8 and prepared by the FBK team[4]. In addition to that, the training data was extended with a Polish - English dictionary, additional (newer) TED talks not included in the original corpus, E-books, proceedings of UK House of Lords, subtitles for movies and TV series, parliament and senate proceedings, Wikipedia Comparable Corpus, Euronews Comparable Corpus and repository of PJIITs diplomas. Also, much of the monolingual data was web crawled from popular web portals and blogs like, chip.pl, Focus newspaper archive, interia.pl, wp.pl, onet.pl, money.pl, Usenet, Termedia, Wordpress web pages, Wprost newspaper archive, Wyborcza newspaper archive, Newsweek newspaper archive, etc. We used linear interpolation and Modified Moore Levis Filtering for in-domain adaptation (Wolk and Marasek, 2014). The pre-processing of the corpora included tokenisation, cleaning, factorisation, conversion to lower case, compound splitting, and a final cleaning after splitting. Tuning was performed for each system.

The baseline system was prepared using the Moses open source SMT toolkit with its Experiment Management System (EMS) (Wołk and Marasek, 2014). The SRI Language Modeling Toolkit (SRILM) (Stolke, 2002) with an interpolated version of the Knesser-Ney discounting (interpolate unk kndiscount) was used for 5-gram language model training. We used the MGIZA++ tool for word and phrase alignment. KenLM (Heafield, 2011a) was used to binarise the language model, with a lexical reordering set to use the msd-bidirectional-fe model (Wołk and Marasek, 2014).

For experiments and training we used Moses SMT with Experiment Management System (EMS) (Wołk and Marasek, 2014). Starting from baseline (BLEU: 16.70) system tests, we raised our score through extending the language model with more data and by interpolating it linearly. We determined that not using lower casing, changing maximum sentence length to 95, maximum phrase length to 6 improves the BLEU score. Additionally we changed the language model order from 5 to 6 and changed the discounting method from Kneser-Ney to Witten-Bell. Those setting proved to increase translation quality for PL-EN language pair in (Wolk and Marasek, 2013). In the training part, we changed the lexicalised reordering method from msd-bidirectional-fe to hier-mslr-bidirectional-fe. The system was also enriched with Operation Sequence Model (OSM) (Durrani et al., 2011). What is more we used Compound Splitting feature (Wołk and Marasek, 2015). Tuning was done using MERT tool with batch-mira feature and n-best list size was changed from 100 to 150. This setting and language models produced the score of BLEU equal to 21.57. Lastly we used all parallel data we were able to obtain. We adapted it using Modified Moore Levis Filtering (Wołk and Marasek, 2015). From our experiments we concluded that best results are obtained when sampling about 150,000 bi-sentences from in-domain corpora and by using filtering after the word alignment. The ratio of data to be kept was set to 0.8 obtaining our best score equal to 23.74.

Even though the speech translation experiment deals only with the PL-EN direction, for the sake of completeness, we report the EN-PL results here as well. Because of a much bigger dictionary, the translation from EN to PL is significantly more complicated. Our baseline system scored 9.95 in BLEU. Similarly to PL-EN direction we determined that not using lower casing, changing maximum sentence length to 85, maximum phrase length to 7 improves the BLEU score. Additionally we set the language model order from 5 to 6 and changed the discounting method from Kneser-Ney to Witten-Bell. In the training part, we changed the lexicalised reordering method from msd-bidirectional-fe to tgttosrc. The system was also enriched with Operation Sequence Model (OSM). What is more, we used Compund Splitting feature and we did punctuation normalisation. Tuning was done using MERT tool with batch-mira feature and n-best list size was changed from 100 to 150. Training a hierarchical phrase-based translation model also improved results in this translation scenario (Graliński et al., 2013). The best score for this direction was 22.76.

To address the issue of rather low BLEU scores, it was quite difficult to obtain any reasonable training material for the translation module, given limited time to perform these tests. That is

[4] https://sites.google.com/site/iwsltevaluation2014/mt-track

Input	BLEU
Reference	20.68
tri2b	8.06
FMLLR	9.58
Capitalised/punctuated tri2b	9.17
Capitalised/punctuated FMLLR	11.00

Table 7: Speech translation results for the baseline subtitles only SMT system.

Input	ASR vocab	BLEU
Reference	n/a	27.35
tri2b	small	10.13
FMLLR	small	12.43
FMLLR	large	20.59
FMLLR oracle	large	22.01

Table 8: Speech translation results based on the improved TED SMT system.

why it has been decided to use a system that is as close to the domain as possible, i.e. the TED lectures. PJIIT also added as much data as possible (as outlined in 1.4.2), but it was expected that the results were going to be poor. In the end, the performance of the SMT system was even slightly better on the PJIIT lecture data (27.35 BLEU) than on the original TED evaluation set (23.74 BLEU). PJIIT continues their efforts to obtaining a sufficient amount of translated lecture material as a basis for a decent speech translation system.

1.4.3 Speech translation experiment

Initial experiments were performed on an SMT system trained using only movie subtitle and their results outlined in Table 7. The performance of that system on the actual reference files of the audio transcripts was around 20 BLEU. The outputs of the two ASR models (using large beam, but smaller vocabulary) were then fed into the system to produce results of 8 and 9.5 BLEU respectively. An attempt was then made to use automatic capitalisation and punctuation of the output, which gave an additional increase in BLEU of about 14% in either case (actual values are in table 7). The automatic capitalisation/punctuation system is still very preliminary, but does show some promise. We suspect that an accurate punctuation/capitalisation method with digit and abbreviation generating facilities would give an even better result, but that is beyond our capabilities at the moment.

In the second run of the experiments, an improved SMT model as described in the section above, was used. The results are outlined in Table 8. The same (raw) input as in the first experiment performed much better but a noticeable difference occurred when the ASR utilised the large, 214k word vocabulary. Nevertheless, there is still a considerable gap between the scores that use ASR output compared to the reference. Even an almost ideal oracle ASR system (only 7 % WER!) lags quite a bit behind the reference system. This demonstrates how the SMT system is quite sensitive to the errors in the ASR, much more than the WER measure would suggest.

2 Task 6.2: External evaluation campaigns

The partners of the EU-BRIDGE project participated in the leading evaluation campaigns in machine translation and speech translation. This involved organising the campaigns and submitting systems to be evaluated. In this way it is assured that the systems developed by the consortium deliver state-of-the art performance.

The consortium also made contributions to the way these campaigns are run. A crucial point is the development of reliable and meaningful ways to rank systems. FBK developed a novel evaluation protocol for IWSLT based on post-editing. HKUST also applied its semantic evaluation protocol to the task. Details can be found in the attached papers at the end of this deliverable.

2.1 International Workshop on Spoken Language Translation (IWSLT)

The IWSLT evaluation campaign in 2014 was to a large degree organised by the EU-BRIDGE project. The overview paper is attached to this deliverable. The academic partners in the project very actively participated in the campaign.

- KIT built speech recognition and machine translation systems (Kilgour et al., 2014; Slawik et al., 2014)
- RWTH built speech recognition and machine translation systems (Wübker et al., 2014)
- FBK built speech recognition and machine translation systems (Babaali et al., 2014; Bertoldi et al., 2014)
- UEDIN built speech recognition and machine translation systems (Bell et al., 2014; Birch et al., 2014)
- PJIIT built speech translation systems (Wolk and Marasek, 2014)
- HKUST built machine translation systems (Beloucif et al., 2014)

There was also a joint speech translation submission by partners of the EU-BRIDGE project (Freitag et al., 2014b) and a joint speech recognition submissions for English.

Details results and how well the participants did in the respective conditions can be found in the evaluation overview paper which is attached to this deliverable.

2.1.1 Cross-fertilisation of technologies and system design within IWSLT

In order to demonstrate how the concept of coopetition as also applied for the IWSLT evaluation campaign drives progress within the research community and also within the consortium, we give below an overview of the development of some salient technology innovations in the last three IWSLT evaluation campaigns. The list shows, how techniques get picked up by additional participants over time and improve the systems across sites, while other techniques might be dropped due to ineffectiveness.

Automatic speech recognition In the last three IWSLT ASR evaluation campaigns the following groups participated.

- RWTH - RWTH Aachen University, Germany
- KIT - Karlsruhe Institute of Technology, Germany
- FBK - Fondazione Bruno Kessler, Italy
- UEDIN - University of Edinburgh, United Kingdom
- MITLL-AFRL - Mass. Institute of Technology/Air Force Research Lab., USA

- LIUM-Vecsy - University of Le Mans and Vecsys, France
- NAIST - Nara Institute of Science and Technology, Japan
- NICT - National Institute of Communications Technology, Japan

For these groups we were able to identify the following techniques that were picked up by more and more participants over the last three evaluation campaigns:

- Subspace GMMs
  - IWSLT 2012: NICT
  - IWSLT 2013: NAIST, NICT, IOIT
  - IWSLT 2014: FBK, NICT
- Neural network language models
  - IWSLT 2012: NICT
  - IWSLT 2013: UEDIN, NICT, IOIT
  - IWSLT 2014: FBK, UEDIN, NICT, MITLL-AFRL
- Topic adaptation
  - IWSLT 2012: NICT, NAIST, FBK
  - IWSLT 2013: FBK, NICT, RWTH,
  - IWSLT 2014: NICT
- System combination
  - IWSLT 2012: NICT, KIT-NAIST, KIT
  - IWSLT 2013: FBK, KIT, NAIST, NICT, RWTH, IOIT, UEDIN
  - IWSLT 2014: FBK, UEDIN, KIT, IOIT, NICT, VECSYS-LIUM
- Lightly supervised AM training:
  - IWSLT 2012: KIT-NAIST, KIT, NICT, UEDIN, FBK
  - IWSLT 2013: UEDIN, KIT, FBK, RWTH, NICT, IOIT
  - IWSLT 2014: FBK, KIT, IOIT, UEDIN, NICT, MITLL-AFRL
- Bottle neck neural network feature extraction / Tandem MLAN:
  - IWSLT 2012: UEDIN
  - IWSLT 2013: KIT, RWTH, UEDIN
  - IWSLT 2014: KIT, UEDIN, IOIT, MITLL-AFRL
- HMM-ANN hybrid acoustic models:
  - IWSLT 2012:
  - IWSLT 2013: KIT, NAIST, NICT, UEDIN
  - IWSLT 2014: FBK, KIT, UEDIN, IOIT, NICT, MITLL-AFRL
- DNN speaker adaptation:
  - IWSLT 2012:
  - IWSLT 2013: NICT
  - IWSLT 2014: NICT

Machine translation In the last three IWSLT MT evaluation campaigns the following groups participated.

- RWTH - RWTH Aachen University, Germany
- KIT - Karlsruhe Institute of Technology, Germany
- FBK - Fondazione Bruno Kessler, Italy
- UEDIN - University of Edinburgh, United Kingdom
- PJIIT - Polish-Japanese Institute of Information Technology, Poland
- MITLL-AFRL - Mass. Institute of Technology/Air Force Research Lab., USA
- QCRI - Qatar Computing Research Institute, Qatar Foundation, Qatar
- LIMSI - France
- LIUM - University of Le Mans, France
- NAIST - Nara Institute of Science and Technology, Japan
- NICT - National Institute of Communications Technology, Japan
- USTC - National Engineering Laboratory of Speech and Lang. Inform. Proc., Univ. of Science and Techn. of China
- USFD - University of Sheffield, United Kingdom
- MSR-FBK - Microsoft Corporation, USA, and FBK
- NTT-NAIST - NTT Communication Science Labs, Japan and NAIST

For these groups we were able to identify the following techniques that were picked up by more and more participants over the last three evaluation campaigns:

- Word Class Language Model Wuebker et al. (2013)
  - IWSLT 2012: RWTH, KIT, MITLL-AFRL
  - IWSLT 2013: RWTH, KIT, UEDIN, MITLL-AFRL
  - IWSLT 2014: RWTH, KIT, UEDIN, MITLL-AFRL, USTC
- Punctuation Prediction with Monolingual SMT System Peitz et al. (2011); Cho et al. (2012)
  - IWSLT 2012: RWTH, UEDIN
  - IWSLT 2013: RWTH, KIT, UEDIN, MSR-FBK
  - IWSLT 2014: RWTH, KIT, UEDIN, FBK, USFD, LIMSI
- KenLM Toolkit Heafield (2011b); Heafield et al. (2013)
  - IWSLT 2012: -
  - IWSLT 2013: KIT, UEDIN, FBK, QCRI, PJIIT
  - IWSLT 2014: RWTH, KIT, UEDIN, FBK, MITLL-AFRL, USFD, PJIIT
- Neural Network Language/Translation Model Niehues and Waibel (2012); Schwenk et al. (2012); Auli et al. (2013); Sundermeyer et al. (2014)

  - IWSLT 2012: KIT, MITLL-AFRL
  - IWSLT 2013: RWTH, KIT, NTT-NAIST, IOIT
  - IWSLT 2014: RWTH, KIT, UEDIN, MITLL-AFRL, NTT-NAIST, LIUM, LIMSI, USTC, NICT

- Improved Domain Adaptation Moore and Lewis (2010); Axelrod et al. (2011); Bisazza et al. (2011)
  - IWSLT 2012: UEDIN, FBK
  - IWSLT 2013: RWTH, KIT, UEDIN, FBK, NTT-NAIST, QCRI, MITLL-AFRL, IOIT
  - IWSLT 2014: RWTH, KIT, UEDIN, FBK, NTT-NAIST, USFD, USTC

- Operation Sequence Model Durrani et al. (2013)
  - IWSLT 2012: -
  - IWSLT 2013: UEDIN, QCRI
  - IWSLT 2014: UEDIN, LIUM, USTC, PJIIT

- TED Data Cettolo et al. (2012)
  - used by all teams in all IWSLT

2.2 ACL Workshop on Statistical Machine Translation (WMT)

The Workshop for Statistical Machine Translation is a series of annual events that concerns itself with text-to-text translation of news with large scale resources (up to one billion words of parallel text, billions of words of monolingual text) for ten language pairs (French, Spanish, German, Czech and Russian into English and back). UEDIN plays a core role in organizing this campaign, which is mainly funded by the EU FP7 CSA MosesCore. The following academic partners in the EU-BRIDGE project participated in the campaign in 2014.

- KIT (Do et al., 2014; Herrmann et al., 2014)
- RWTH (Peitz et al., 2014)
- UEDIN (Durrani et al., 2014; Williams et al., 2014)

There was also a joint submission by partners of the EU-BRIDGE project (Freitag et al., 2014a).

3 Significance of results

For all automatic scores of the speech recognition evaluations in our internal campaings and the external IWSLT evaluation campaign and and for machine translation evaluation in IWSLT we performed tests for assessing the statistical significance of differences between the participating systems.

For the automatic speech recognition systems we used the sc_stats tool of the NIST SCTk scoring suite. With the tool we performed the McNemar test at utterance error level (MN), the Matched Pairs Sentence Segment Word Error Test (MP), the sign test at speaker level (SI) and the Wilcoxon Signed Rank Test at speaker error level (WI). For all tests we used a 95% confidence interval. By default sc_stats computes those scores pairwise between all systems that ran on a specific test set.

For the machine translation systems automatic scores were computed using multeval by Johnathan Clark Clark et al. (2011) which implements a stratified approximate randomization. Due to the larger number of systems and the larger number of conditions in IWSLT for machine translation, only the significance of every system compared to the best performing system was computed.

The detailed output of the significance tests can be found in Appendix A, while below we give a summary of the findings.

3.1 Euronews evaluation

For the following languages two separate submission were received and their differences were tested for statistical significance. All details can be found in Appendix A.1

Arabic The MN test claimed significance for all language pairs, while all other tests failed.

English All systems were significantly different, only the MP test for KIT and FBK failed.

Italian The KIT system is significantly different from the FBK and PEV systems, while for the FBK and PEV pair the MP and WI test failed.

Polish The PJIIT and RWTH systems are significantly different, except for the SI test.

Portuguese The PEV system is significantly different from the RWTH and KIT systems. The RWTH and KIT system are not different according to the MP and WI test.

Russian The FBK and KIT system are significantly different according to all tests.

Turkish The FBK and KIT system are not different according to the MP and WI test.

3.2 Sky News evaluation

For the Sky News evaluation for all system pairs the MN test failed. Otherwise, the systems from FBK, KIT and RWTH were not significantly different. Further the system from PEV was otherwise significantly different from all other systems. Also, the PEV and KIT systems were significantly different in the remaining tests from the UEDIN system, while for the pair RWTH and UEDIN only the MP test showed statistical significance. For the pair FBK and UEDIN the SI test failed in addition to the MN test, while the other two tests claimed statistical significance. Detailed results can be found in Appendix A.2

3.3 IWSLT ASR evaluation

Here we summarise the findings of the significance tests for all ASR submissions to IWSLT 2014. Detailed results can be found in Appendix A.3.

English For the IWSLT ASR evaluation in English all significance test showed statistical significance except for the following system pairs and tests:

1. FBK and KIT: MP, SI, and WI failed
2. FBK and LIUM: SI and WI failed
3. KIT and LIUM: SI and WI failed
4. LIUM and UEDIN: MP, SI and WI failed

German For the IWSLT ASR evaluation in German all tests claim statistical significance for all system pairs.

Italian For the IWSLT ASR evaluation Italian the following tests failed for the following system pairs:

1. FBK and KIT: SI and WI failed
2. FBK and MITLL-AFRL: MP, SI, and WI failed
3. KIT and MITLL

3.4 IWSLT MT evaluation

All systems were significantly different from the best performing system, except for the UEDIN system in the English-German SLT evaluation and the LIUM and RWTH systems in the English-French SLT evaluation. Details can be found in Appendix A.4.

4 Task 6.3: Field testing

4.1 Speech translation support within the European Parliament

The European Parliament is particularly interested in a tool which might help the interpreters with their preparation work and also to find key terms easily during debates. For this purpose the interpreter support tool (soon coined 'IST' by EP staff) has been developed by KIT and released at `https://www.interpreter-support.eu`. The interpreters can upload files and can access the service of terminology extraction, together with terminology translation, and named entity tagging on the uploaded files and on-line European Parliament documents. To showcase sufficient coverage and at the same time keep engineering effort reasonable, we currently limit the IST to English as a source language and, regarding the full functionality, to five target languages (German, French, Spanish, Polish and Finnish), all specified by the EP.

Two rounds of field testing have been carried out. The first round of field testing took place from September 29th 2014 to October 27th 2014. Six EP interpreters had volunteered to participate in the test and gave feedback. The second round was held from January 12th 2015 to January 20th 2015. Eighteen EP interpreters participated in this field test. The interpreters used the tool in their real preparation for a work assignment using EP documents. 76% of the volunteers used the tool from time to time, while 29% just used it for this field testing.

The interpreters were asked both to provide feedback in the form of free-text and a fill-in questionnaire after each test round. The questionnaire included 13 sections which cover the

general impression on the tool as well as the opinions regarding its helpfulness, the interface, the quality of service, etc. The questionnaire is accessible at `http://www.smartsurvey.co.uk/s/134763SFIBY`.

After discussions during bilateral meetings and a presentation of the concept, as well as its initial implementation, the first round of the field test was the first opportunity for KIT to collect feedback of participating interpreters who had really used the system, and to improve the IST system based on this information. Both the free-text feedback and the responses to the standardised questions from the final round, and their comparison with the first results show that the system has substantially improved. Figure 1 shows the general impression on the IST tool. More than 60% of the interpreters are satisfied or very satisfied with the final tool. This figure almost doubles the indicative value of the first round (34%). It should be noted here

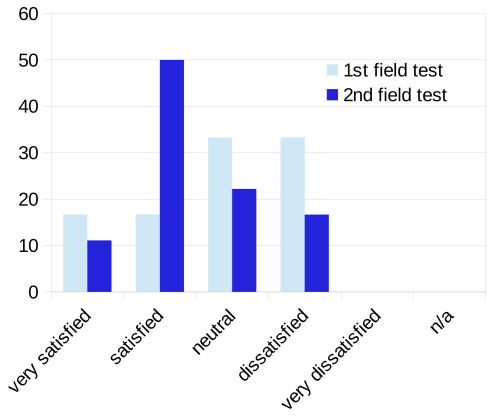


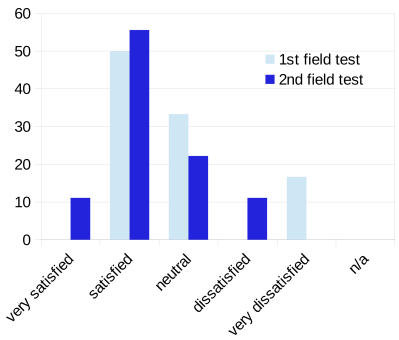
Figure 1: Satisfaction with the interpreter support tool in general.

that the number of volunteers is (by construction) larger in the second than in the first round, 18 instead of 6. Thus, the results of the second round are more reliable. For the first round, it was sufficient to get a rough indication in which direction the system had to be improved. The number of field testers in the second round actually exceeded our expectations and again indicated the high interest from the EP side.

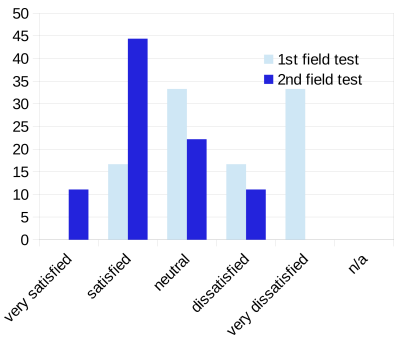
Figure 2 shows a more detailed analysis regarding different aspects of the tool, notably (a) the tool's interface, (b) the service of terminology extraction, (c) the terminology translation service, and (d) the named-entity extraction. The non-negative feedback (very satisfied, satisfied or neutral) on each of the four aspects exceeds 60%. The positive feedback (very satisfied or satisfied) on ease of use and terminology extraction exceeds 60%. Summarising Figure 1 and Figure 2, most of the interpreters are satisfied with the interpreter support tool, especially with the interface and the terminology extraction.

The feedback on the IST tool's design is shown in the Figure 3. In the implementation much effort has been put into the design of the IST tool, in order to make it easy and user-friendly for the interpreters, to deal with both the European Parliament committees and plenary documents. Eight aspects of it have been explored in the questionnaire. Figure 3 indicates that, for the final round of the field test, most parts of the tool are satisfying. Almost 90% of the interpreters have considered this tool pleasant to work with (Figure 3(a)) and around 80% of volunteers have agreed that the design of the interface is attractive (Figure 3(b)). Figure 3(c), Figure 3(d), and Figure 3(e) refer to the file management, where the interpreters can import EP documents from the EU website or upload their own files. About 80%-90% interpreters have given positive feedback on the functions. However, only 40% interpreters have stated that this tool has all the functions that they need (Figure 3(f)). Such an opinion suggests that in the future this tool should be extended with new functions to fully satisfy the needs of interpreters.

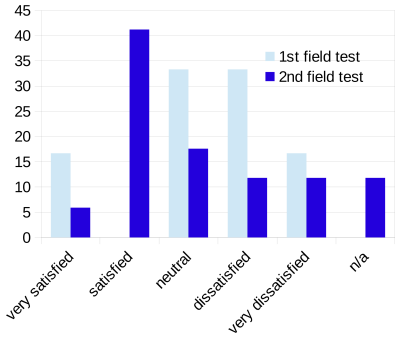
The results of the quality evaluation are shown in Figure 4. The positive feedback on overall quality, terminology extraction and named-entity extraction is over 50% and on the translation



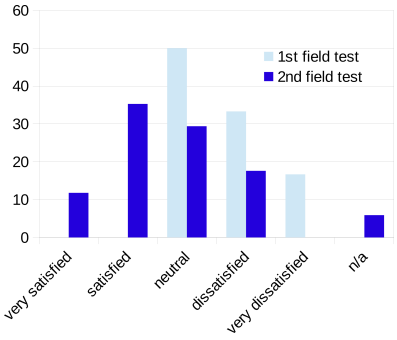
(a) Satisfaction with ease of use of IST.



(b) Satisfaction with terminology extraction.



(c) Satisfaction with the translation service.



(d) Satisfaction with the NE extraction.

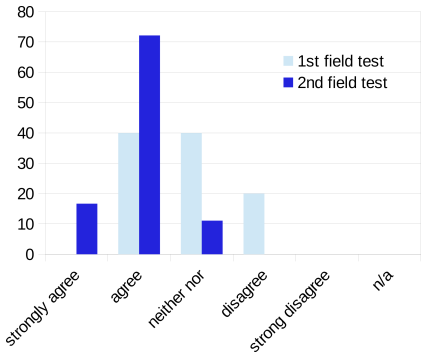
Figure 2: Degrees of satisfaction with IST in general and with each service.

service a little below (43.8%). The latter does not refer to MT aspects but to the access to two on-line reference lexicons. Furthermore, the comments of the interpreters indicate that they typically find the tool very promising and appreciate what it can already offer. At the same time, they expect some improvements to follow. On the basis of the test we learnt that the next expected improvements on terminology extraction are to reduce the number of noisy terms, to miss fewer terms, to reduce the number of common terms, and to show only the base form. On the translation service side, the following aspects should be improved: too many noisy translations, too many translation missing, some translations are not accurate or are not satisfactory for certain languages. Regarding the named-entity extraction, the following aspects should be improved: too many noisy terms, too many terms missing, too many named entity types in the number category.

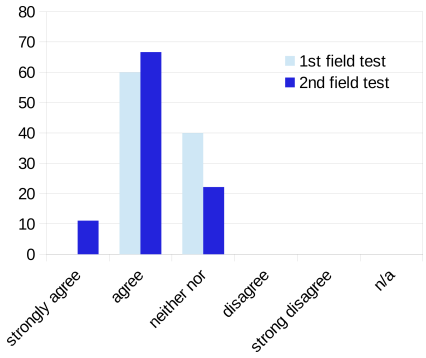
Another important evaluation aspect is the helpfulness of the IST tool for the interpreters. The results are shown in Figure 5. The positive feedback on the overall helpfulness, terminology extraction and named-entity extraction is for each aspect over 50%.

On the translation service for the terminology terms, the positive feedback is only 39% and 27.8% for n/a. A comparatively large fraction has not used the term translation feature - it might as well be that they did not need it on the documents under consideration, but it probably indicates that it is not perceived as being useful. This result could be attributed to a couple of reasons. First, the translation service is influenced by the quality of terminology extraction: When the terminology string is wrong, it is hard to find a corresponding translation in another language. Second, the translation experience is hampered by the already mentioned lack of coverage by the on-line lexicons. And third, in their proper field of expertise, interpreters have high demands, while they might expect less from the NE tagger or the terminology extraction which just provide a speed-up when working through a document.

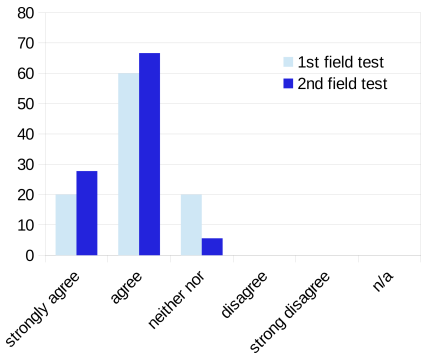
Bearing in mind that the interpreter support tool can and should be further improved, it is fair to say that its field test was highly successful. The testers of the system, many of them not particularly accustomed to IT, and all of them highly skilled and demanding when it comes to linguistic and translation aspects, provided rather positive feedback on the tool, its functions



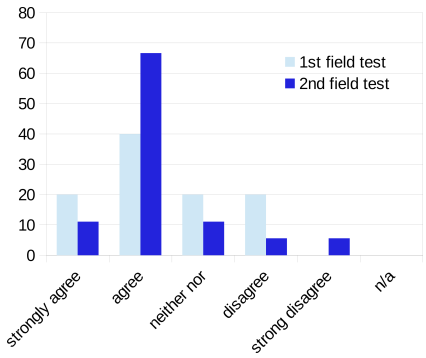
(a) The tool is pleasant to work with.



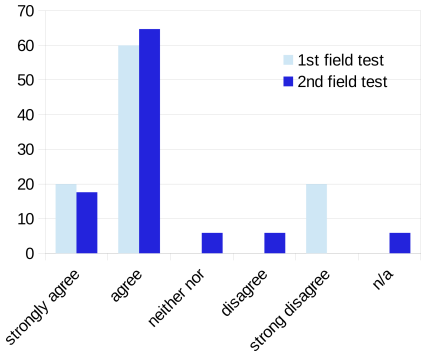
(b) The design of the tool is attractive.



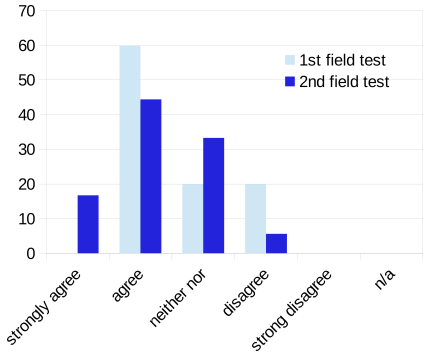
(c) It is easy to find EP documents.



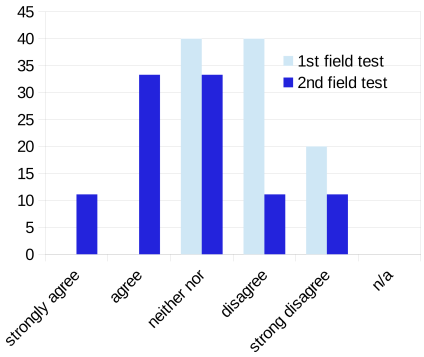
(d) The document management is simple.



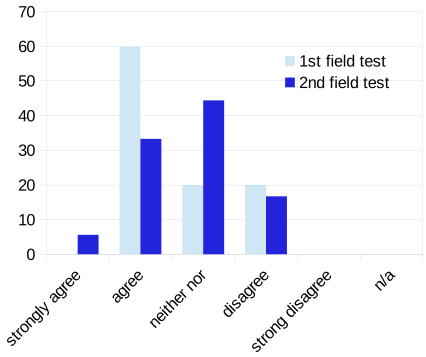
(e) Uploading documents works well.



(f) The navigation within the tool is clear.

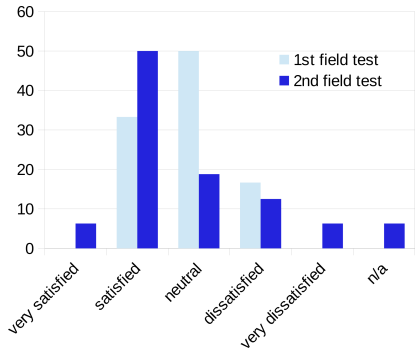


(g) The tool has all the functions I need.

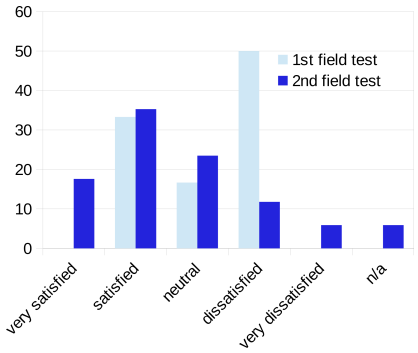


(h) All features are explained well.

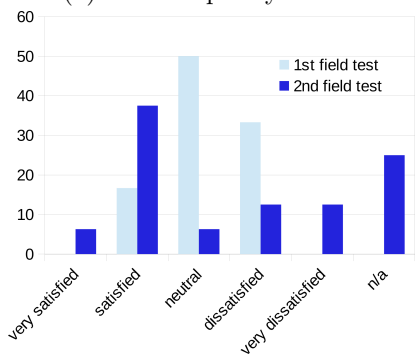
Figure 3: Degrees of agreement with statements on the interface.



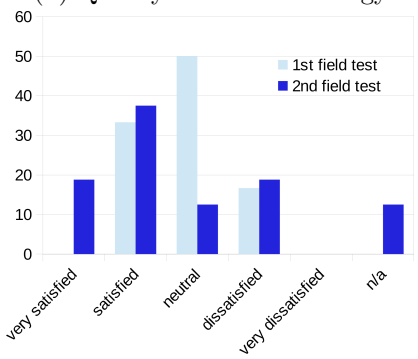
(a) Overall quality of the IST system



(b) Quality of the terminology extraction.

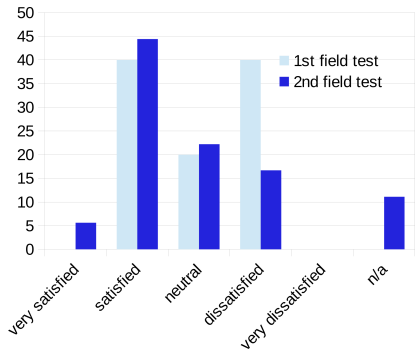


(c) Quality of the translation service.

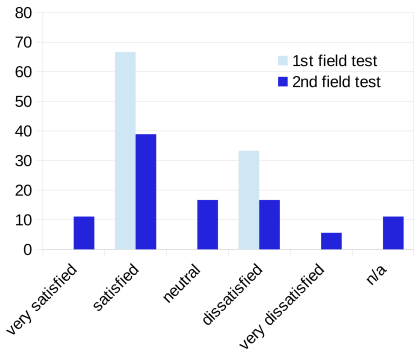


(d) Quality of the named-entity extraction.

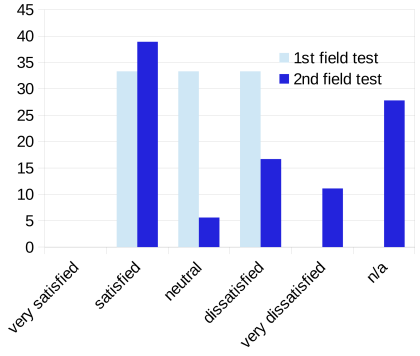
Figure 4: Quality evaluation overall and per service.



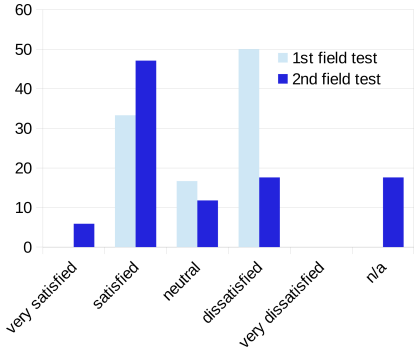
(a) The overall helpfulness of the IST tool.



(b) The helpfulness of the term. extraction.



(c) The helpfulness of the translation service.



(d) The helpfulness of the NE extraction.

Figure 5: Helpfulness of the IST tool in general and per service.

and components. Thus, we can conclude that the system is cracking an important barrier by offering a solution good enough to be appreciated and to be found helpful by the target users. This view is reinforced by the fact that our EP contacts have clearly articulated their interest in a continuation and they are looking for internal funds to pursue the improvement and further use of the tool.

We believe that this is also a first step to encourage the EP authorities, including the interpreters and their management, to pay more attention to language technology and speech translation technology services that might be offered to the European Parliament in the future.

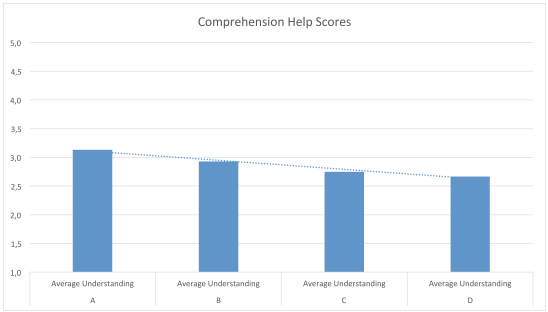
4.2 Unified Communication translation service

The 8 kHz vs. 16 Khz question (see P3 report) brought some new constraints on the unified communication field test.

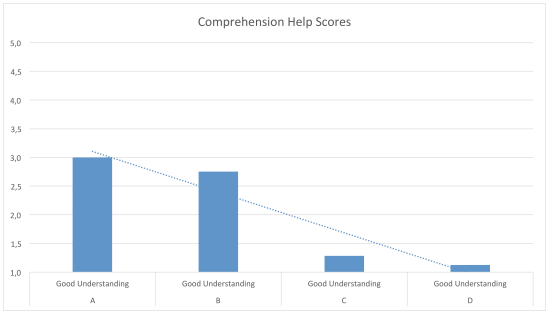
For the field test, in order to be successful and representative of what Serenty will provide in a few months, we had to come up with an auxiliary platform that would provide/simulate a 16 kHz unified communication platform.

Player. This platform, called the player, has been developed only for the sake of supporting the field test (see P3 Periodic Report).

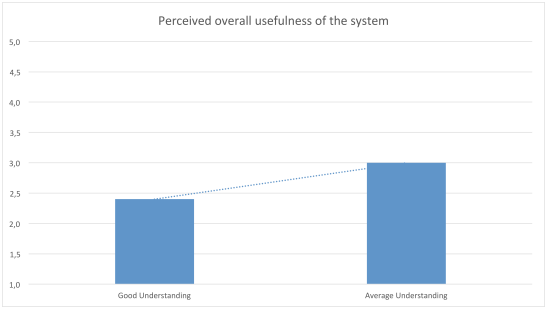
The rationale behind the player was to decouple webinar recording from webinar following. For this a system has been built that would simulate transcription and translation in real-time (producing the transcription and the translation at the rhythm of our decoders) while allowing presenters to do their job independently of the time zone they are with respect to that of the French test users. This simplified the organisation of the field test as this way we did not have to organise series of meetings for when all French test users and the webinar would take place at the same time.



(a) Overall helpfulness.



(b) Helpfulness of the terminology extraction.



(a) Helpfulness of the translation service.

4.2.1 The Webinars

The webinar field test was run in Paris with 10 French students listening to 12 [5] webinars presented in English by 5 Americans and 4 non-native English speaking Europeans totalling 4 hours and 39 minutes of speech. The subjects of the webinar where mainly business and marketing related, whereas one covered a physics theme and another one an IT theme. However, these two technical webinars were done at a very high level.

The webinars can be viewed at the following link: `http://rct.uniquity-web.andrexen.com/serenty-play/`.

The content of each webinar is summarised in Table 9. It is organised per speaker indicating the domain that the webinar or the series of webinars belongs to, a short title as well as a reference number ("Web. #").

Domain	Title	Webinar summary	Web.#	Speaker
Business	Business Strategies	What are the different strategies a company can follow and how to differentiate: an introduction in 3 short webinars	4,9,1	David
Marketing	Big Data value	Mintigo presents the way information is extracted on internet for enterprise customers	5,2	Mintigo
Marketing	Digital marketing	Marqui shows what is important in digital marketing	3	Marqui
IT	WebRTC + Tweet in self service	Two webinars on 2 technical concepts used in the customer service market: Tweet content analysis and WebRTC	6,8	Tobias
Physics	Invisible reality	Visualisation of quantic physics: comparison of the atome (invisible reality) with a baseball (visible reality)	7	Brian
Business	Introducing new technologies	What is needed for the successful introduction of new technologies on the example of the civil aviation and the making of Sauce Maltaise	10	Christian
MT	BLEU scoring	How to evaluate the quality of automatic translation: presentation of the BLEU method	11	Volker
Business	Entrepreneurship	First steps for entrepreneurs	13	Franz

Table 9: List of webinars presented during the field test

Each webinar was linked to a dedicated worker that used a language model adapted to its content (slides, website). The player platform of Andrexen presented the video or slide show of each webinar together with the English transcription and the French translation of the input signal. The transcription and the translation where presented to the test users in real time with exactly the delays given by the system, typically one to two seconds for the real-time decoding and a sentence (as defined by the transcription engine) for the translation.

[5] A total of 13 webinars have been recorded and one webinar has been left aside as a reserve

4.2.2 The users

The users were all French test users in the age between 18 and 25 years old, 5 females and 5 males. The webinars were in English and were automatically transcribed and translated into French. The users have been asked to qualify their proficiency of the English language (in written, speaking, reading, listening). This allowed us to build two categories of users, a group of 5 persons having a good proficiency in English (2 females and 3 males) and a group having an average proficiency in English (3 females and 2 males, see Table 10).

English fluency perception	Gender: female	Gender: male	(total)
Average understanding	3	2	5
Good understanding	2	3	5
(total)	5	5	10

Table 10: Users English proficiency and gender distribution

The test users have been asked to view about 3:15 hours of webinars from the total of 4:39 hours available. We have built four thematic groups: two thematic groups were containing pure business themes, named Bus1 and Bus2 and the other two thematic groups contained a mix of business and IT themes, which were called BusIT1 and BusIT2. Test users have chosen to follow one thematic group according to their interest. The thematic groups are presented in Table 11 .

Bus1				Bus1		
webinar #	Speaker	Time		webinar #	Speaker	Time
4,9,1	david	0:32:44		4,9,1	david	0:32:44
3	marqui	0:40:32		5,2	mintigo	0:39:49
6,8	tobias	0:59:32		6,8	tobias	0:59:32
10	christian	0:42:21		10	christian	0:42:21
7	Brian	0:21:20		7	brian	0:21:20
Totals	8 webinars	3:16:29		Totals	9 webinars	3:15:46
Bus-IT1				Bus-IT2		
webinar #	Speakers	Time		webinar #	BusIT1	Time
4,9,1	david	0:32:44		4,9,1	david	0:32:44
6,8	tobias	0:59:32		6,8	tobias	0:59:32
3	marqui	0:40:32		5,2	mintigo	0:39:49
11	volker	0:26:26		11	volker	0:26:26
12	franz	0:16:13		12	franz	0:16:13
7	brian	0:21:20		7	brian	0:21:20
Totals	9 webinars	3:16:47		Totals	10 webinars	3:16:04

Table 11: The 4 different webinar thematic groups test users could choose from

4.2.3 Offline Evaluation

The major goal of the offline evaluation is to calculate the WERs (word error rates) obtained on the different webinars. The WER calculated will give a basis for the analysis on usefulness sense on the side of the users and see if there is a correlation between this usefulness sense and WERs.

The second goal of the offline evaluation is to formerly analyse the impact of the adaptation methods developed for this use case.

It was not expected that adaptation of the language model would improve WER dramatically, not to say in a significant manner (from 40.8% to 40.6% as can be seen in Table 12 is not significant at all). But for the sake of a better user experience, we were expecting that words that have been identified as new and important for the webinar but that would without adaptation be out of vocabulary (OOV) should be recovered so they could be recognised and translated by the system. And indeed, in average over all evaluated webinars, 44% of the OOVs have been recovered.

Webinar	Native	Topic	Unadapt. WER	Adapted WER	OOVs recov.	OOVs non recov.	% OOV recov.	Unnecessary new words	Grade (A-D)
David.Altern	Yes	Bus	15.7%	15.3%	0	7	0%	8	A
David.Intro	Yes	Bus	8.74%	8.7%	4	3	57%	6	A
David.whatis	Yes	Bus	13.6%	11.9%	0	6	0%	8	A
Marqui.Market	Yes	Mark	50.5%	51.7%	27	35	44%	74	C
Brian.Quantics	Yes	Phys	22.9%	22.9%	16	76	17%	103	A
Mintigo.Intro	Yes	Mark							B
Mintigo.Data	Yes	Mark							B
Volker.BLEU	No	Tech	64.9%	-	0	0	0%	0	D
Christian.Market	No	Bus							D
Franz.Bus	No	Bus							D
Tobias.Tweet	No	Tech	53.1%	52.9%	163	126	56%	594	C
Tobias.WebRTC	No	Tech	36.7%	34.9%	22	46	32%	597	C
Totals			40.8%	40.6%	232	299	44%		

Table 12: WERs on each of the webinars with adapted and non-adapted workers. Bus == Business, Mark == Marketing, Phys == Physics, Tech = Technological nature. (For grade see online evaluation)

4.2.4 User tests

Each student had to fill a form for each of the webinars they viewed as well as a form at the end of the field test in order to give an overall feedback on their experience with the unified communication platform (see Appendix

In order to draw some correlation between notes given by test users to their perception on viewing each webinar with the platform and the quality of the transcription/translation provided by the system, we have grouped the 12 webinars into 4 distinct categories depending on quality. The grouping is a function of WER, if available. For the one third of the seminars that was not transcribed (and hence no WER available), we made the classification based on clearness of speech and accent using our professional experience (and, needless to say, prior to the user tests). The 4 groups are called grade A to grade D as described in Table 13.

Standardised test of comprehension-help perception First we are looking at how transcription has been perceived by test users relatively to the grade of the webinar. And we have split test users in 2 groups, those thinking having a good proficiency in English (good

Webinar-Grade	A	B	C	D
Word Error Rates	10-25%	30-50%	40-60%	Above 60%
Clearness of speech	Good	Normal	Normal	Normal
Native vs non-native speaker	Native	Native	Non-native, light accent	Non-native, strong accent

Table 13: Webinar grade as a function of WER, clearness and accent

understanding), and those thinking having an average proficiency (average understanding) in English.[6]

One important question in the feedback form was about the help the system can provide test users with understanding better each webinar. What was not surprising is the drop in "comprehension help" noted by test users having a good proficiency in English while going from grade B to grade C webinars (see Figure 8) [7], dropping the grade given on "the system helps me understand the webinar better" by 1.5 MOS points. This means that for these test users (having a good English proficiency), there was a clear cut for webinars of grade C: for these test users, the quality of the transcription and translation was too bad to be considered helpful.[8]

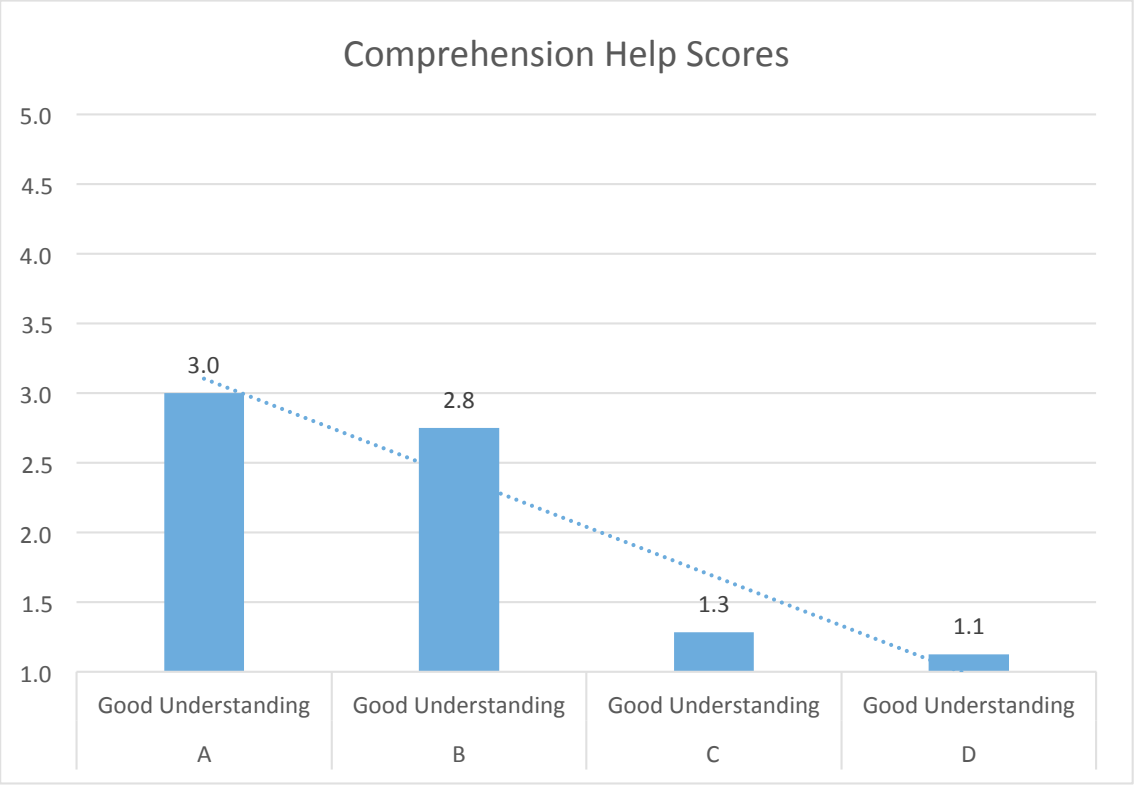


Figure 8: Comprehension help perceived by test users having a good understanding of English for A-grade to D-grade webinars

But what is very interesting is that this drop in "comprehension help" is not observed at all with test users having only an average proficiency (average understanding) in English (as opposed to those having a good understanding of English). Even for grade D webinars (see Figure 9). It seems test users with average English proficiency still found words or expressions

[6]While we do not have much data (10 students listening to 4.5 hours of content) from which we could draw hard conclusions, we do see some indications on how our system has been perceived by the students.

[7]Points given by the students are between 1 (worse) and 5 (best) where 3 is an average grade.

[8]Points given by the students are between 1 (worse) and 5 (best) where 3 is an average grade.

they are missing even in the transcriptions of webinars of grade C or D where the WERs are rather bad. This is an indication that important words are still well recognised and that their appearance in the transcript is perceived useful, even if the sentence is not necessarily readable.

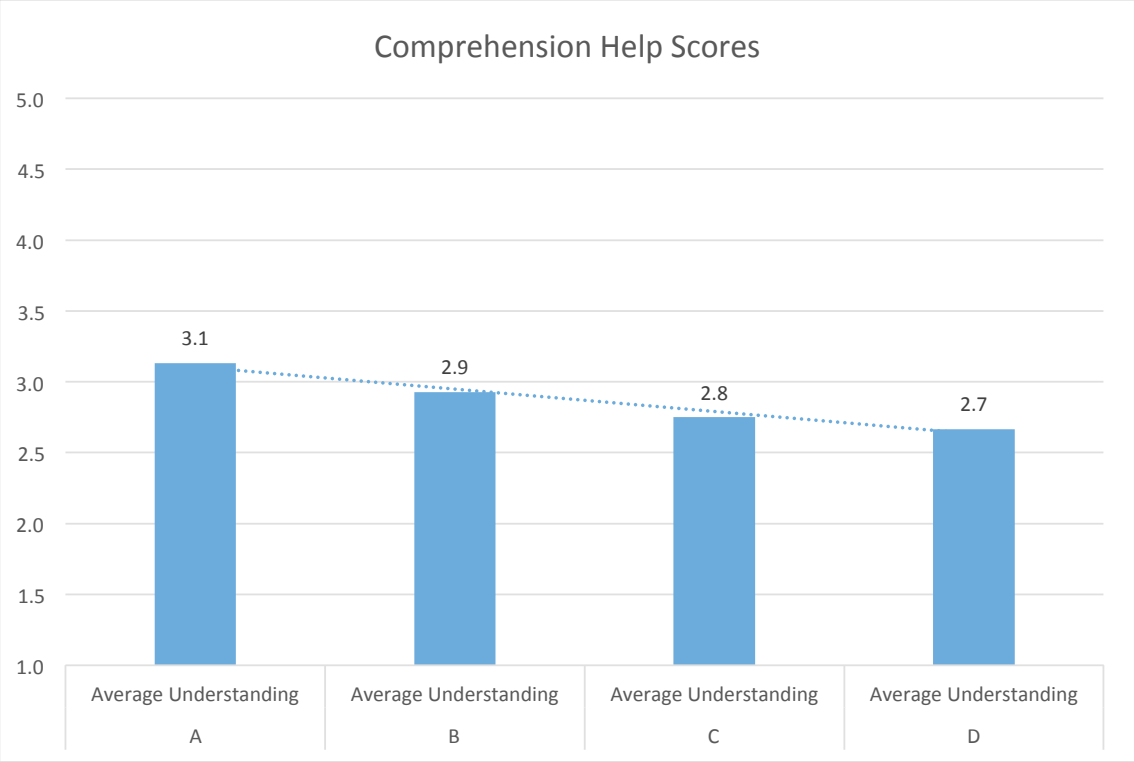


Figure 9: Comprehension help perceived by test users having an average understanding of English for A-grade to D-grade webinars

Is the system useful? In the end, the main question is to state whether the system is useful or not.

After having seen all the webinars (from grade A to grade D), the test users with average proficiency in English, consider the system somewhat useful (see Figure 10 on the global perception given by test users), with an average score of 3, a score that is higher than the one given by test users with good proficiency (score 2.4) [9].

Free-form feedback from test users Test users have been asked to make their own comments on their experience using the system. While quantification of comments is not possible, comments give a very good insight in the overall impression left over by a system to the testers. When grades given on the feedback forms are average and do not allow to draw much conclusions, it is extremely helpful to revert to the written free-form feedback. If overall comments given by testers were negative, the system would indeed be bad, but if overall comments given by testers were constructive, the system would be perceived as very interesting even if a few things were missing.

And we read plenty positive and encouraging overall comments from the test users. They even give ideas on how to improve the system by proposing features (e.g. highlighting keywords). Table 14 gives an extract of comments (removing duplicates) made by the test users.

[9] This score is given by the students at the end of the field test, after having viewed good and not so good transcriptions and translations. Hence the average scores (covering all webinars) do not reflect the postiveness of the comments made in a free text form (probably strongly influenced by the last webinar which has been well perceived)

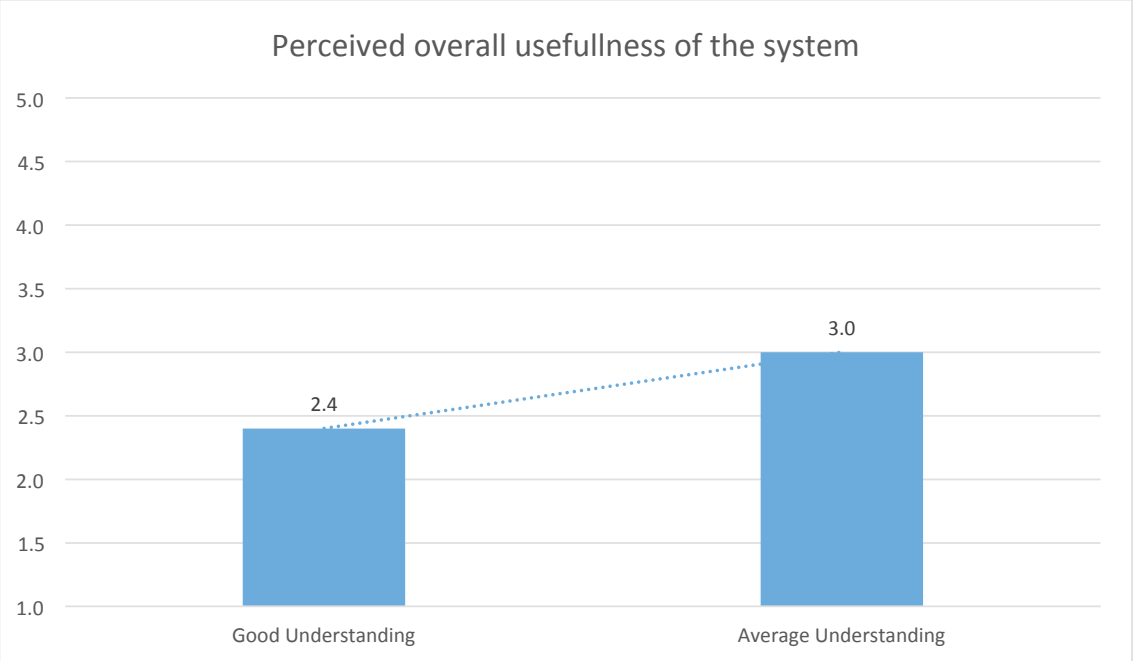


Figure 10: Global perception on the system for test users having a good understanding of English vs. test users having an average understanding of English

Original comment (with typos)	**English translation of comment**	**How often it appeared**
Continuez	Continue	1
On arrte pas le progrs!	Technical progress never stops!	1
Bientt la transcription instantane!	Instant transcription is soon here	1
Ce serait intressant de pouvoir revenir en arrire (scroll back) dans la traduction/transcription la main pour aller vrifier un terme par exemple.	Scroll back function would allow to verify a word	3
Diminuer la latence entre la video et le texte transcript	Reduce latency between video and transcription	4
Mais proposer des keywords serait un bon compromis !	Propose (highlight) keywords would be a great compromise	3
personnellement, j'aurais prfr n'avoir que la transcription anglaise mais mon niveau d'anglais est trs bon...	I would have need only the English transcription as my English is very good	1
Beaucoup d'usage de sigles la place de mots courants dans la transcription	Acronyms appear too often in place of normal words	1
Pas mal de problmes avec les dates (et heures?) exemple : "nineteen sixty-three" retranscrit comme "19" "...six. T-shirt"	Lots of problems with date and time, Example: nineteen sixty-three transcribed as 19 six. T-shirt	1

Table 14: Overall comments and suggestions made by test users on their impression while using the system

4.2.5 Conclusion

The very positive general comments the test subjects have made at the end of the field test are very encouraging.

The users also came with ideas on how to improve presentation of the transcribed and translated content. For example that keywords like product names and/or acronyms should be highlighted. Indeed, enriching the text (both in source and target language) e.g. with named entity information, such as in the European Parliament interpreter support system, promises to provide additional value to the webinar translation.

The adaptation process using the content of each webinar has proven successful as in average 44% of all OOVs have been recovered, i.e. they have been correctly recognised after adaptation. Which is as important for the presenter, who wants to be sure his/her very specific topic is well transported (e.g. the system tries to understand his/her specific jargon), as for the listeners as these are often very relevant content words.

An important conclusion is that the feeling of usefulness for a transcription and translation system given by the test subjects does not stop when the transcription makes the written text difficult to understand (e.g. with WER above 25%). The tests indicate that even for webinars with a WER as high as 50%, listeners with average proficiency still found in the transcription and translation some of the words they unable to translate and hence felt that the system was helpful. Generally speaking, the tests indicate that listeners with an average (rather than good) understanding of the webinar source language i.e. the ones who really depend on the system seem to tolerate an even weak system performance more than we had anticipated.

Therefore we may want to augment the figure on understandability zone from the P3 report with an additional bar on perception of usefulness of the system for users with average English proficiency (see Figure 11) keeping the understandability zone as the usefulness perception zone for users with good English proficiency.

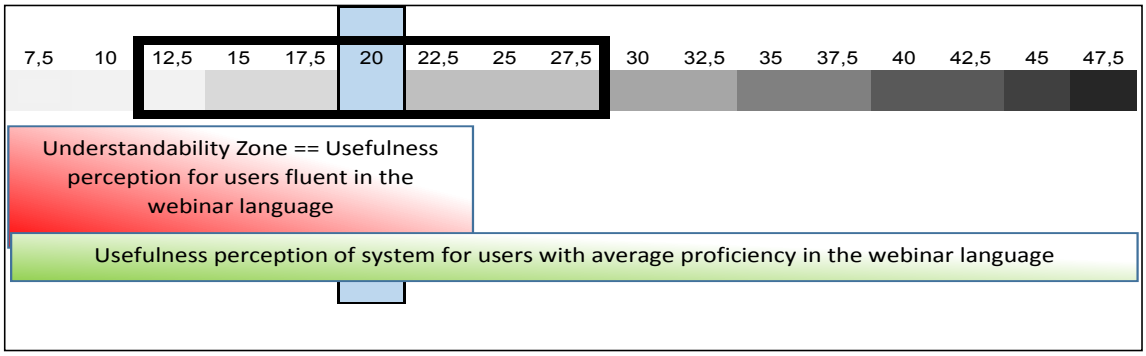


Figure 11: Relation between WER, understandability (red bar) and usefulness (green bar)

Based on these comments, we are encouraged to enhance the system highlighting named entities in a sentence. We may even think about using the ASR scores and define for example two score-thresholds: a word or phrase scored above the higher threshold could be written in a larger font than the words or phrases scored between the two thresholds, and presenting in a faded grey those words or phrases scored below the lowest threshold.

As a conclusion, Table 15 and Table 16 below show our understanding of the market value and market potential of a transcription and translation system of webinars for different configurations.

The cells marked "Yes" show that in this configuration, as much listeners with good proficiency as listeners with average proficiency in the language of the speaker will find the system

useful. This is the sweet spot market to address first in order to gain experience and improve the system with. This market shall be addressed within the next 6 months.

The cells marked (Yes) show that in this configuration, listeners with average proficiency in the language of the speaker will find the system useful. This market should be carefully monitored.

In all other cases ("Later"), the system should be introduced in the market later on, after having gained sufficient experience with needs and after having improved the two underlying technologies.

Mode	Unilateral Presentation mode				Conversational mode with channel separation			
Speaker Type	Professional speaker		Non-professional speaker		Professional speaker		Non-professional speaker	
Signal quality	8kHz	16kHz	8kHz	16kHz	8kHz	16kHz	8kHz	16kHz
Native Speaker	(Yes)	Yes	Later	(Yes)	Later	(Yes)	Later	(Yes)
Non-Native Speaker	Later	(Yes)	Later	Later	Later	Later	Later	Later

Table 15: Sweet spots for a real-time unified communication transcription and translation system

Table 15 summarises our understanding for a real-time system, so transcribing and translating spoken content on the spot and hence with a small delay between sound and text.

As this time delay between sound and text appearance (especially that of translation which has an inherent longer delay than that of transcription) has been perceived by test subjects as annoying, we may think of a market opportunity for content that is already recorded and available online. In this case, we could deliver the written content as it is spoken (no delay or even in advance of time or available all at once and high-lighting the current words) and also with a higher quality as more CPU time and 2-pass decoding would be available to achieve a better WER (see Table 16).

Mode	Unilateral Presentation mode				Conversational mode with channel separation			
Speaker Type	Professional speaker		Non-professional speaker		Professional speaker		Non-professional speaker	
Signal quality	8kHz	16kHz	8kHz	16kHz	8kHz	16kHz	8kHz	16kHz
Native Speaker	(Yes)	Yes	Later	Yes	Later	(Yes)	Later	(Yes)
Non-Native Speaker	Later	Yes	Later	Later	Later	Later	Later	Later

Table 16: Sweet spots for a non-real-time unified communication transcription and translation system

4.3 Caption translation services for television broadcast

4.3.1 Outline of field tests

Field testing on Caption Translation Services began in October 2014. The purpose of the field tests was to determine two things:

Deployment Whether EU-BRIDGE technologies could be successfully and reliably deployed by Red Bee Media to produce television captions for Weatherview and Sky News.

Accuracy To what extent any resulting captions were of an accuracy standard suitable for broadcast according to UK regulatory requirements.

It was clear from the beginning that the accuracy of any automatic system would be outperformed by experienced and professional captioners. However, it was important to use the official accuracy metrics in order to quantify the gap and as a basis for communication with Red Bee management.

While the consortium had originally planned to run the field tests only on Weatherview, the results there were so promising that Red Bee asked technology partners, at relatively short notice, to add a second target of Sky News broadcasts.

4.3.2 Overview: MCloud for captioning

The EU-BRIDGE technology being field tested was a subtitling service realised via the MCloud mediator, designed and maintained by PerVoice, in combination with the workers from the consortium partners.

MCloud provides a C library which implements the raw XML protocol used by the Service Architecture and exposes a simplified API for the development of client/workers. For convenience, the library integrates some high-level features like audio-encoding support and data package management. MCloud allows for the use of different audio formats (e.g. PCM, OPUS, SPEEX) according to the available bandwidth and the desired audio quality. It also provides a set of call-back functions used to transform the MCloud output format into a preferred format (e.g. SRT, CTM or TTML).

Since the most commonly-used operating system for desktop environments is Windows, and as most of our applications are based on the .NET framework, we have developed a .NET wrapper of the MCloud API in order to support the design of client desktop applications for the service architecture. This wrapper—called NetMCloud—is available to integrate third-party components, written using the .NET framework, into the service architecture.

The following describes how the client and worker interact in a captioning scenario:

Worker:

- Connects to the mediator, which is running on a specific host and port;
- Specifies a description for each exposed captioning service.
  - A captioning service description is specified by an input stream type, an input fingerprint, an output stream type and an output fingerprint;
- Waits until a new captioning request is received;
- As soon as the new service request has been accepted, every audio packet coming from the client is asynchronously processed by the data callback function. This function contains the captioning logic that is applied to the audio stream. The result is sent back to the client as an MCloudWordTokens array;
- When the captioning service ends, a complete status message is sent to the client.

Client:

- Connects to the mediator, which is running on a specific host and port;
- Announces a captioning request.
  - A captioning request is specified by an input stream type, an input fingerprint, an output stream type and an output fingerprint;
- Waits for a worker to accept its request
- As soon as this is done, audio packets can be sent;
- Data packets coming from the worker are asynchronously processed by the data callback function. This function contains the logic that is applied to the MCloudWordTokens coming from the worker;
- When all audio packets are sent to the worker, a complete message is sent to the worker.

For the first three sets of field tests, Mediator V1 was used. This was updated to Mediator V2 for the field tests on 7th January 2015 and thereafter. The second mediator version is an extension of the first, adding the following features:

- An accounting logic, in order to store tickets about workers utilization. Currently this information is used in order to detect the status of a worker (busy or idle);
- An extension of the path generation algorithm for enforcing the selection of paths with an exact matching of the input and output fingerprint requested by clients;
- A software component evolution, intended to increase maintainability and performance.

4.3.3 Field testing process

Field testing of MCloud for captioning services was undertaken by two members of Red Bee Media staff. One focused on testing the deployment of MCloud and Red Bee Medias ability to access MCloud. The other focused on the accuracy of the resulting captions.

The deployment tests were run on nominated days. Partners engaging in the field tests were notified of these days in advance and asked to ensure their workers are available for testing.

The following bulletpoints outline the process followed by Red Bee when undertaking field tests:

- Each worker (or worker pair if there is a punctuation option) is tested once with 'unseen' audio. This audio is either approximately 3 minutes of Weatherview, including a verbal introduction over music, or approximately 10 minutes of Sky News, beginning with the musical channel sting and headlines voiceover. All workers receive the same test clips.
- If no output is received then another attempt will be made later in the day after first verifying that the worker is available via the MCloud Service Monitor webservice.
- If incomplete output is received, this will be submitted for marking if it appears there were no technical issues other than the worker stopping text submission. Otherwise, another attempt will be made as above.

- If an ASR worker is not available on the MCloud interface, another attempt will be made later the same day. If it is still not available, there is no test and the worker will be noted as unavailable.

- If a Punctuation worker is not available then the test will be run using only the ASR worker (i.e. unpunctuated text will be submitted for accuracy assessment).

- If MCloud fails, testing will be rescheduled.

Tests were run on a single day and repeated multiple times over the course of several weeks. The intention was to use the results to address potential deployment issues and to track accuracy improvements as engines continued to be developed during field testing. Field test results were uploaded to the EU-BRIDGE intranet for dissemination to the project partners.

4.3.4 Further information on data used in testing

Tests were undertaken on samples of Sky News and BBC Weatherview data. A total of 17 engines were expected to be made available by partners over the course of the field tests. Of those, seven were Weatherview-trained engines and 10 were Sky News-trained engines. Some partners submitted multiple engines for testing, with the aim of producing punctuated text as well as unpunctuated text.

The test sample was the same for all partners on test days. Participating partners were instructed to make their engines available and avoid any development work on published test days. The sample was taken from that days broadcast output. As such, no partner could have worked on the sample output prior to field testing.

4.3.5 Metrics used for accuracy measurement

Red Bee Media uses an internal mark scheme to determine the accuracy of captions. This scheme was applied to captions produced by MCloud in order to assess their accuracy. Sky News content was marked from the top of the hour to 10:00 minutes on audio file. For Weatherview, only actual weather content marked, so continuity announcements either side ignored. An audio transcript was also created by the reviewer for each sample. There are three kinds of error were marked:

(Missing words) 1 mark per missing word

Recognition errors 1 mark per incorrect word

Style errors 1 mark per error

Missing words Words which are in the audio but the engine did skip during recognition. One error counted for each missing word.

Audio:	Some heavy, thundery rain possible across Wales.
Text:	**for** heavy thundery **may support** across **swales West**

Audio:	Showers for Thursday perhaps	
Text:	showers Thursday perhaps	here one error is marked for the missing word "for"
Text:	showers **forty** perhaps	one error counted for misrecognition

Audio:	a sunnier start to tomorrow
Text:	a **sunny** start to tomorrow

Incorrect words Words which appear in the text but to not match audio. One error counted for each incorrect word. Anything which looks like code or machine instructions has been treated as incorrect words:

<SIL> <noise> the *<noise> <SIL>*

Any word or words which match the audio are marked as correct, even if the surrounding text makes them meaningless:

Here the word across would be of no use to a viewer, but matches the audio so is counted as correct.

In some instances there is potential overlap between incorrect and missing words. Here only the incorrect words are marked.

Any deviation from the audio should be marked as an error, even if the sentence still makes sense.

Style errors Marks are deducted for formatting and capitalisation errors- "BBC **one**", "**northern Ireland**", "a **Sunny** start to the day" etc.

Missing/extra capital letters should be marked as one error per incident, rather than one error for each word. So "**the lion, the witch and the wardrobe**" would only count as one error. Any incorrect words within that also count as errors.

Numbers are marked as correct regardless of whether they are rendered as words or digits. Errors have not been marked for style around compass points- "in the south/South", "northwest/north-west" etc.

Punctuation Punctuation workers were provided by some partners alongside unpunctuated workers. As textual accuracy can be assessed without punctuation, we took a standard position not to mark missing punctuation as incorrect. The one exception was where apostrophes are missing in a word but present elsewhere in the sample and error is counted- e.g. "it ll" rather than "itll".

From a commercial perspective, results from punctuated engines would be much closer to the desired outcome of the use case, and good results from punctuated engines would be preferred to good results from unpunctuated engines. However, the scope for assessing accuracy development in unpunctuated workers encouraged Red Bee to include them in the field testing, on the assumption that development work on punctuation would continue and be enhanced by strong textual accuracy.

4.3.6 Calculating caption accuracy

Percentage score is calculated by:

$$\#Words_{transcript} - (\#Words_{incorrect} + \#Words_{missing} + \#Errors_{style}) = \#Words_{correct}$$

$$\frac{\#Words_{correct}}{\#Words_{transcript}} \times 100 = \text{Percentage correct words}$$

A separate score is provided that discounts style errors, as these are seen as the least problematic and most easily fixed of the three.

Sky News Marked from the top of the hour to 10:00 minutes on audio file.

Weatherview (also known as "Weather For The Week Ahead") Only actual weather content marked, so continuity announcements either side ignored.

An audio transcript is created by the reviewer for each sample. The runtime for each test run is being measured from the transmission of the first audio packet until the reception of the last text packet.

4.3.7 Results

I. Service handling and reliability As there was a sufficient number of test days, the partners did not execute a dry run before the field test. As such, the first one or two of the five test dates were also used to identify configuration issues.

After resolving firewall issues and other minor concerns, the handling of the service (simple APIs) was good. The infrastructure worked well but Red Bee personnel experienced a service reliability that did not meet the high standards of a broadcast environment. In the results table, it is clearly marked whether a result was available on first-time request or whether calling the service had to be repeated (under otherwise identical conditions, as outlined above). To give an idea of the type of issues encountered and resolved during field testing, please see the following examples and explanations:

- 5 out of 7 engines successfully returned appropriate results for Weatherview, 6 out of 8 for Sky News (Field test 1, 29 Oct 2014). (The original test shows two additional Sky News workers which turned out to just be backup systems identical to two other workers.)
- KIT engine stops delivering results (temporary disk full)
- RWTH engine not visible (time-out issue)
- KIT engine stops delivering results (software revision issue)
- Mediator not available (failed software upgrade)
- FBK Sky News engine returns text in German (wrong configuration)

All of these issues could be individually traced and resolved but it can be seen that they all relate to similar instances of localised failure. The conclusion is that, for a professional-grade highly-reliable service, the experimental system development and the operational service architecture should be strictly separated such that a stable service on dedicated servers with dedicated personnel and with rigid software revision management can be guaranteed. While EU-BRIDGE has demonstrated the feasibility and provided valuable insights, it also underlines the need for a separated commercial activity; even more so if the service is going to scale up by several orders of magnitude. Based on EU-BRIDGE 's achievements, a professional European service infrastructure should be set up.

II. Quality of the service architecture The infrastructure worked well on the whole; reliability issues were related to the local environment and have been discussed above. The interface and client tools developed were greatly improved by feedback and co-operation between the various partners; it is true to say that the success of the field testing process was down to the unified interfaces between all parties and that any future large consortium project should consider such an approach to avoid unnecessary effort being wasted on establishing interfaces between engines and engine users.

III. Quality of the transcripts For use in a commercial setting recognition accuracy (including punctuation) is key. The performance of a speech recognition system depends to a great extend on the acoustic conditions of the recording. In this setting, recordings featured broadcast professionals being recorded using professional equipment. Although the acoustic conditions are very good, there is a challenge because of added artificial noise—music, background noises and chatter—which features particularly heavily in live news broadcasts.

In this section, we present results of the field test. Numbers in the tables that are marked with an asterisk * refer to tests that needed a re-run due to technical or other issues. Bold numbers indicate the best worker in that condition for that day. "n/a" means that no test ran for that condition on that day. We tested workers with and without punctuation. The ones with punctuation are marked "Punct.". In general, the recognition accuracy varies between test runs of different days. This is due to the different properties of the individual audio files used for testing.

The results from the Weatherview field test are shown in Table 17. It can be seen, that the workers from RWTH and UEDIN produce output with the highest accuracy, while the workers from PerVoice (PEV) show the lowest score.

Worker	29.10.	12.11.	26.11.	7.1.	14.1.
KIT	n/a	n/a	84.5%*	89.2%*	90.1%
RWTH	**97.8%**	90.9%	n/a	**95.7%***	**90.3%***
UEDIN	n/a	**91.9%**	**95.1%**	93.9%	89.0%
PEV	79.3%	73.9%	82.4%	80.2%	70.3%
KIT Punct.	n/a	n/a	84.7%*	89.2%*	**87.1%***
UEDIN Punct.	**96.3%**	n/a	n/a	n/a	78.7%

Table 17: Accuracy scores from Weatherview field test, bold numbers indicate the best result for each day.

Table 18 shows the results from the Sky News evaluation. Like in the Weatherview use-case, RWTH produced outputs with the highest accuracy. Unlike for Weatherview, the UEDIN workers produced mediocre results.

Worker	29.10.	12.11.	26.11.	7.1.	14.1.
KIT	n/a	n/a	74.8%	n/a	n/a
FBK	n/a	n/a	n/a	76.0%*	71.1%*
RWTH	**87.7%**	**90.6%**	**88.8%**	**90.1%**	**83.0%**
UEDIN	n/a	68.0%	63.6%	n/a	81.1%
PEV	64.8%	76.7%	n/a	74.8%	65.8%
KIT Punct.	n/a	n/a	**74.9%**	**79.0%***	**74.9%**
UEDIN Punct.	**64.2%**	**62.6%**	62.8%	n/a	n/a

Table 18: Accuracy scores from Sky News field test, bold numbers indicate the best result for each day.

The last table, Table 19, in this section shows the results from the Sky News Online field test. These workers are designed to produce output in real-time. This is achieved at the cost of quality.

Worker	7.1.	14.1.
UEDIN	74.6%	**79.3%**
PEV	**79.3%**	71.1%
KIT Punct.	78.7%*	n/a

Table 19: Accuracy scores from Sky News Online field test, bold numbers indicate the best result for each day.

IV. Runtime of workers In order to assess the real-time factor of the different workers, we measured the run-time of selected test runs and converted the run-time into real-time factors. The results are shown in Table 20 for Weatherview and in Table 21 for Sky News.

Worker	7.1.	14.1.
KIT	**0.47**	**0.97**
RWTH	0.51	1.02
KIT Punct.	0.47	0.99

Table 20: Real-time factors of different workers for Weatherview, bold numbers indicate the best result for each day.

Worker	7.1.	14.1.
RWTH	**0.25**	**0.25**
FBK	1.23	1.35
KIT Punct.	0.37	n/a

Table 21: Real-time factors of different workers for Sky News, bold numbers indicate the best result for each day.

V. Business case There are two business cases in connection with this use case. One is an efficiency improvement for pre-recorded captioning by streamlining manual processes for producing prepared captions, for which 100% accuracy is demanded. The second is a risk management solution for live captions, providing a real-time on-air ASR service to cover captioner drop-out, for which 98% accuracy is demanded. This would avoid potential fines from clients due to the loss of airtime.

The results were convincing enough to let Ericsson (Red Bee's mother company) create a commercial follow up; see "4. Final Remarks" in D5.2.3

4.4 BBC

An EU-BRIDGE team participated in the BBC #newsHACK-III event in London on 15-16 December 2014, organised by BBC Connected Studio, BBC News Labs and BBC World Service (`http://newshack.co.uk/newshack-iii-language-tech/`). The theme of the event was Automated International News Services of the Future, with a focus on Language technology — speech recognition, speech synthesis, machine translation, and information extraction. EU-BRIDGE partners field tested ASR and MT technology developed in the project, by developing prototype systems using APIs developed in EU-BRIDGE . The Edinburgh EU-BRIDGE prototype system, *GlobalVox*, won the prize for "Best end to end multi-language tools".

4.5 Lecture Translator

In order to evaluate the Lecture Translator (LT) system in the field, KIT set itselves the following questions: Do students using the LT understand the lecture better? Does the LT help students? Do students like the interface and can they handle it with ease?

KIT thus decided on five evaluation methods to have their questions answered and to get a global and comprehensive view of the actual situation: (1) Automatic measurements in form of web access and duration of stay on the website, (2) individual interviews, (3) user observation, (4) short surveys, and (5) a comprehensive questionnaire.

4.5.1 Time line / overview

The first part of the field test of the lecture translator system was conducted during the summer term 2014. Automatic metrics, e.g. the number of people using the LT and the average duration of use of the LT per session per person were collected throughout the term. Small surveys were sent around every 3–4 weeks to users that had manifested interest for the project. Towards the end of the term, KIT conducted personal interviews, a survey among students, and observed two students while they were using the system.

For the second part of the field test, conducted in the winter term 2014/2015, KIT strengthened its communication effort (cf Section 4.5.2) and offered the service in more lecture halls. The second part of the field test was carried out towards the end of the term, more specifically in calendar week three (12th to 16th January). Automatically collected data was once more gathered throughout the term. Two short surveys have been sent around.

4.5.2 Communication

At the beginning of each term, the system was presented to the students. KIT staff went to one of the first lectures of each course to give a short talk about the background of the system, to explain how to log in and use it, and to show screen shots. In the summer term, KIT additionally distributed fact sheets with the most relevant information. However, KIT learnt that some students still had not heard about the LT. For the winter term, KIT therefore designed business cards, posters and fliers to improve visibility. KIT also created a new landing page with an easier to read web address as well as a new layout for the web page. Moreover, it asked professors to publish the information concerning the service with other information about their lectures on-line. Posters were put up throughout the university, especially in public places like the university restaurant, in front of lecture halls and at the faculty of informatics. Fliers and business cards were printed and handed to the international office. Business cards were also distributed during the presentation of the lecture translator as a reminder of the web page and the service. Moreover, information was spread via social media and various mailing lists.

4.5.3 Deployment of the system

During the summer term, the system was made available in the Audimax, the main lecture hall of KIT. There, it is fully integrated into the audio system. Moreover, it was installed in one lecture room at the Institute for Anthropomatics and Roboticsb. As the consent of the lecturers is needed to record the lecturers and make the service available to students, KIT was able to run the lecture translator in the lectures shown in Table 22. After the successful implementation in the summer term, KIT made the system available to more students. Thus, KIT installed the system also in the multimedia lecture halls of the computer science faculty. It offered the service to 68 lecturers that were listed for lectures in the on-line lecture catalogue and got a positive answer from 19 of them, teaching 16 different courses. A complete list of recorded lectures is shown in Table 23. In the winter term, more students were interested in the project, 65 (only 25 in summer) put their names down in the list to get regular updates.

Product development - methods of product development
Production management and marketing / production operations management
Higher mathematics
Power-train systems technology A - automotive systems
Basics of computer science
Algorithms I
Computer organisation
Cognitive systems

Table 22: List of recorded lectures during the summer term

Tutorial mechanical design
Programming
Measurements and control systems
Accounting
Higher mathematics I
Higher mathematics III
Finance and accounting
Automatic visual examination and image processing
Automatic speech recognition
Concepts and application of work flow systems
Mechano-informatics and robotics

Table 23: List of recorded lectures during the winter term

4.5.4 Evaluation procedure

Frequency of use - automatic evaluation With the help of the automatic evaluation, KIT wanted to measure:

- the average time of use of the LT per person
- the number of people using the LT per session

Procedure In order to learn more about the number of people using the LT per session, the average time one person stayed logged in and the number of students using the LT during one session, KIT collected anonymised usage data. However, for an exact documentation of the time students stayed logged in, they needed to close the web page. As this was not always the case, the duration of use sometimes stays unknown.

Results For most lectures, the highest activity was registered when the system was initially presented to students in the various courses at the beginning of the term. So students were definitely curious about the project.

In the summer term, there were activities in all seven lectures. In three lectures, however, only the respective day of the presentation of the system and for a short time (about ten minutes). In two lectures (Product development and Computer organisation), there were four activities each, with an average of 2.4 users per session and an average stay of about two minutes. Relatively high activity was observed in the lecture Cognitive Systems, with an average of ten users per session and an average stay per user per session varying from two to fifty minutes.

This was probably due to the fact that the lecture was held by the Interactive Systems Labs and students were not only curious but also regularly reminded of the system.

In the winter term, the highest levels of activity were again recorded the day of the presentation of the system. Between three and 40 students per lecture accessed the site in the different courses. In the fields of computer science, mathematics and economic sciences, KIT registered six lectures in which ten or more people showed an initial interest. During the term, 29 more events (log-in with a duration of stay over one minute) were registered. In five lectures, 2.9 persons in average were active for two or three more sessions. The only exception was once more the lecture of the Interactive Systems Labs, where activities of one to four persons were noted on a regular basis and throughout the term (twelve more times after the initial presentation).

The automated generation of anonymous usage logs offered an insight into how often and to which extend the LT was actually used. While a regular number of users suggests that the LT is appreciated and helpful, the opposite statement cannot be made, as it is impossible to say whether KIT actually reached its target group. German students do not need the LT and probably only have a look at it once, out of curiosity.

4.5.5 Frequency of use - observation of students

By observing students, KIT wanted to get a closer look at how students use the LT and to what extent they are using it.

Procedure Student part timers were instructed to observe the behaviour of students using the LT, focusing on their direction of sight and taking notes on their behaviour. Two students accepted to be observed. The student part timers sat down slightly above and at an angle behind the participants in order to be able to follow their sight as closely as possible and to be able to tell in which direction they were looking: at the screen, the board, the lecturer, or elsewhere.

Results Unfortunately, the data gathered through this experiment was not useful for our task. First of all, it was extremely difficult to find participants, as most students did not want to be observed during a lecture. Second, the participants knew they were being observed and thus probably adapted their behaviour and looked at the screen/LT more often. Third, in most cases it was not possible to say where the students were looking. One used a screen with two or more open windows and it was impossible to tell whether he looked at the LT or at the window next to it. To determine this more accurately, an eye-tracker or a similar device would be needed.

4.5.6 User feedback - exit polls / short surveys

For technical reasons, no exit poll could be integrated into the system. In order to still keep track of the development of the system and detect positive or negative evolutions, anonymous short surveys were sent around on a regular basis (every three to four weeks) via email, each one containing four questions. As the main goal was short feedback, no pre-test was taken.

How often did you use the LT during the last month?
Did you encounter difficulties, if so which ones?
Do you consider the LT useful?
Do you have ideas for useful features or is there anything else you want to tell us?

Table 24: Questions asked in the summer term

Did you use the LT in the last two/four weeks?
What was your first impression of the LT? (November)
How do you like the new features? (December)
Are you going to continue using the LT?
Are there suggestions, ideas or problems you would like us to know about?

Table 25: Questions asked in the winter term

Procedure A link to an on-line survey in English was sent to students that had manifested interest for the LT by signing a list we distributed during the presentation of the lecture translator at the beginning of each term and in the various lectures the service was offered. In summer, 24 students from four different lectures put their names down, 15 of them being foreign students. In winter, 66 students showed interest, with only 21 of them being of German nationality.

However, the answer rates were rather low. In summer, KIT received four answers for the first survey (May) and three for the second (June), in winter KIT had four answers for the first survey (November) and only one for the second (December). In July, respectively in January, KIT ran a larger evaluation and therefore did not send around any short survey. The questions are shown in tables 24 and 25

Results In the summer term, four out of seven students used the LT three times or more in the preceding month, which, considering the fact that most students probably only had one lecture per week in the Audimax with the LT running, is quite good. One used it twice, two used it once, although they did not encounter any difficulties when using the LT. Five students considered the LT useful, one did not answer that question. All in all, especially the transcript was considered helpful. Asked for difficulties, they reported the time lag of the LT, difficulties when logging in, the fact that it is hard to follow the slides and the LT at the same time and the fact that not all lessons were available. A list of suggestions is shown in table 26. In the winter term, KIT also received four answers for the first survey, but only one for the second. Although this does not seem much, the answers were really helpful. Three of the four students actually used the LT. The one who had not used it explained that he understood German and therefore

Chat function that allows foreign students to communicate and help each other when they have problems understanding or when they want to improve the translations.
A lecture archive which would also be useful for German students. They could revisit the lectures along with the transcriptions in multiple languages.
A download possibility for the script at the end of the lecture.
A possibility to show the slides simultaneously as a reference point, thus also avoiding having to alternate between the transcript and the projected screen.

Table 26: Suggestions to improve the system in the summer term

Improve the speech recognition.
A side by side representation of transcription and translation.
A visual aid that helps to keep track of the most recent position.
An indication of the estimated delay between lecturer, recognition results and translated output.
An easier and smoother access to the courses.

Table 27: Suggestions to improve the system in the winter term

did not need the LT. Interestingly enough, he said he would have loved to use the system one year earlier. This entails that it is especially useful for foreign students that start studying in Germany. Opinions on the quality of the LT were mixed. While one student thought that the speech recognition worked rather badly, two described it as good or very good. One person also mentioned the machine translation which he/she considered "very good". Two said they were going to continue using the LT, one wanted to try it from time to time, and one person said he/she would not use it again. One person noted that the delay of the system improved over the time. The suggestions made are shown in Table 27.

In the most recent survey (December), KIT especially asked for feedback concerning the new features. The lecture countdown was considered very helpful, the selection of the lecture hall was not commented on. One student underlined the improvement of the system during the first two months of the winter term. Especially the open questions in the end were helpful, as they provided us with some good ideas that will be taken into consideration in future work.

4.5.7 User feedback - questionnaire

In order to get a more standardised and detailed feedback, KIT designed a comprehensive questionnaire with questions concerning the background of the users, a system evaluation, an evaluation of the components "speech to text transcription" and "machine translation" and a possibility to express ideas and identify problems.

Procedure The questionnaire covered three A4 pages and was distributed to all students that claimed to have used the LT at least once in the summer and/or in the winter term in the frame of the general lecture evaluation conducted by KIT. A small scale pre-test taken within the Interactive Systems Lab's work group made sure that all questions were clear and could be answered within 15 minutes.

In order to increase participation in the winter term and to reach as many students as possible, KIT created an additional on-line survey with the very same questions and asked the lecturers to publish the link with their lecture notes. Thus, even foreign students that work from home were able to answer the survey. KIT also sent the link to all students on their list and published it on the LT-homepage.

All questions where a rating was involved provided a scale ranging from one (worst option) to five (best option) and an additional field n/a, for those cases where the question could not be answered or an answer could not be given. An excerpt of the questionnaire is shown in Figure 12. For the evaluation of the answers, KIT considered three groups, the overall group, the German speaking students, and the foreign students. However, as the sample was small, KIT did not apply any additional statistical actions. Nevertheless, KIT used a weighted average, taking into account all n/a answers.

Results This section discusses the evaluation results from the questionnaire in detail.

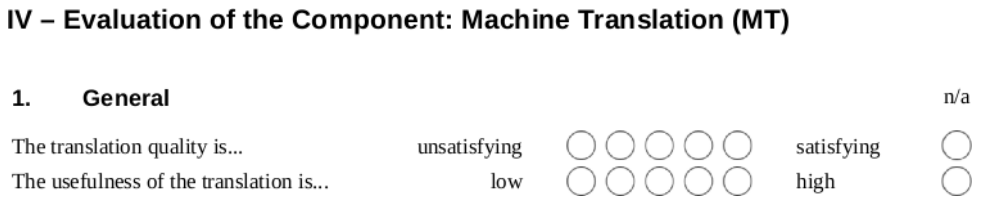


Figure 12: Excerpt from the questionnaire

General Information	All	NG	G
Male	20	4	16
Female	2	1	1
Years of studying in Germany			
less than 1	5	2	3
1-2	7	2	5
3-4	3	1	2
more than 4	5	0	5
n/a	2	0	2
How often did you use the LT?			
1	10	2	8
2-5	10	3	7
6-10	2	0	2
Would you use the LT again?			
yes	14	3	11
no	8	2	6

Table 28: General information, **NG** non-German, **G** German

Part 1 - General Overall, 22 students from five lectures (Computer organisation, Cognitive systems, Programming, Finance and accounting, Accounting) answered the questionnaire, two of them having Chinese as mother tongue, two of them Spanish, one Russian. The level of German of the foreign students was quite good, varying from B2 to C2. Their English level was a bit lower (B2 to C1). All participants had a high level of English, ranking from B1 to C2.

As shown in Table 28, ten students only used the LT once, ten students two to five times, and two students six to ten times. The majority of the participants of the study would use the system again.

The students were mostly male and studied business engineering or computer science. They were studying in Germany between less than one year and more than 4 years.

Part 2 - overall system evaluation The results for the section overall system evaluation are shown in Table 29. The general impression was rather positive, with 3.21 points on a scale from one to five. German students rated the system slightly better than foreign students. It was also considered rather useful, with 3.23 of 5 points.

When asked in more detail about the perceived usefulness, especially foreign students thought that it improved their understanding of the lectures and said they would find it useful in other lectures, too. However, they were not so sure about their performance and whether the LT made it easier to follow the lectures. Some students explained the latter phenomena by saying they sometimes considered it difficult to switch between the lecturer, the slides, and the LT-screen.

The ease of use was also rated positively, with 3.27 points. The layout of the user interface was considered very clear and got the highest marks from both groups.

		All	NG	G
General Impression		3.21	3.00	3.27
The service is...	terrible - wonderful	3.23	3.20	3.24
The experience is...	frustrating - satisfying	3.18	2.80	3.29
The system is...	not useful - useful	3.23	3.00	3.29
Perceived usefulness		3.26	2.8	3.47
Using the LT improves my performance in studying for this subject.	disagree - agree	3.13	2.20	3.55
Using the LT increases my understanding of the lecture.	disagree - agree	3.13	3.20	3.09
Using the LT makes it easier to follow the lecture.	disagree - agree	2.81	2.00	3.18
I would find the LT useful in other lectures.	disagree - agree	3.94	3.80	4.00
Perceived ease of use		3.27	2.65	3.45
I enjoy using the LT.	disagree - agree	3.19	2.00	3.56
The service works as expected.	disagree - agree	2.73	2.00	2.94
The features provided are sufficient.	disagree - agree	2.86	2.20	3.06
The layout of the user interface is clear.	disagree - agree	4.27	4.40	4.24

Table 29: Overall system evaluation, **NG** non-German, **G** German

Part 3 - Speech-to-text (STT) component The evaluation results of the STT component are shown in Table 30. Generally, the impression of this component was good, especially among foreign students who found the transcription very useful. The overall quality was evaluated with 3.01 points. The largest difference of opinion in this section was observed in the category of transcription errors. Those were considered less distracting by foreign students. The lowest mark in this section got the delay of the transcription (2.62 points), the highest mark the transcription of general terms (3.44). Foreign students considered the quality of technical terms even better than German students.

The section "usefulness" got an overall mark of 3.13. Foreign students said STT had helped to improve their performance in the subject and their comprehension of the lecture. Only some agreed that the transcript made it easier to follow the lecture.

Part 4 - Machine Translation (MT) Component In part four, students were asked to evaluate the machine translation component. The results are presented in Table 31. The general impression was even better than the impression of the STT component (3.19 points MT vs. 2.95 STT). Foreign students, however, appreciated the STT component more (3.4 STT vs. 2.67 MT). Although the translation quality was considered quite high, its usefulness was rated a bit lower, especially by foreign students.

When asked in more detail about MT quality, the results were similar to the answers to the question in the first section. Foreign students generally rated the STT component higher than the MT component. The lowest mark among foreign students was given to the delay in translation, which was considered rather high. They also found the translation quality fluctuating. Nevertheless, students observed an improvement of the quality over the term. The scores of the quality of translation of general terms were rated highest in this section, by both groups.

		All	NG	G
General		2.95	3.40	2.82
The transcription quality is...	unsatisfying - satisfying	2.86	3.40	2.71
The usefulness of the transcription is...	low - high	3.05	3.40	2.94
Quality		3.01	3.13	2.97
The errors of the transcription were...	distracting - not distracting	2.75	3.40	2.53
The delay in transcription was...	high - low	2.62	2.80	2.56
The transcription was...	disfluent - fluent	2.90	2.60	3.00
During lectures, transcription quality was...	fluctuating - consistent	3.06	3.20	3.00
During the term, transcription improved...	not at all - clearly	3.15	3.00	3.20
The transcription of general terms was...	bad - good	3.44	3.40	3.46
The transcription of technical terms was...	bad - good	3.28	3.50	3.21
Usefulness		3.13	2.93	3.22
The transcription helped me improve my performance in studying for this subject.	disagree - agree	3.19	3.00	3.27
The transcription made it easier to follow the lecture.	disagree - agree	3.06	2.80	3.18
The transcription made it easier to comprehend the content of the lecture.	disagree - agree	3.13	3.00	3.20

Table 30: Evaluation STT component, **NG** non-German, **G** German

Interestingly enough, the amount of questions rated with n/a was quite high in this part. More than 30% of the statements were not rated. Six people, including one foreign student, did not express any opinion about MT quality at all. This was similar for the section on usefulness, where nearly 40% of all possible answers was rated n/a, and eight people did not express their opinion on the usefulness of MT at all. The remaining students considered the usefulness of the MT component a bit lower than the usefulness of the STT component (2.83 MT vs 3.13 STT). In this section, the differences between German speaking and foreign students were rather distinct, showing a difference of more than 1 point. A possible explanation is that, as the lectures were in German, the STT component added value to the lecture, whereas the English output was an additional language to process and thus made it more difficult.

Part 5 - Ideas and problems The last part asked for suggestions and things to improve. Ten students responded by putting in one or more comments. Suggestions are shown in Table 32. Things to improve are shown in Table 33. There were also comments that showed that students found the project interesting, that they were interested in its future development and above all its further improvements.

4.5.8 User feedback - interview

By conducting personal interviews based on a standardised questions, KIT wanted to get more information about what users really think about the LT. Interviews also allowed to ask for more

		All	NG	G
General		3.19	2.67	3.36
The translation quality is...	unsatisfying - satisfying	3.20	3.00	3.27
The usefulness of the translation is...	low - high	3.18	2.25	3.46
Quality		3.13	2.68	3.28
The errors of the translation were...	distracting - not distracting	2.86	2.50	3.00
The delay in translation was...	high - low	2.87	2.25	3.09
The translation was...	disfluent - fluent	3.00	2.50	3.17
During a lecture, translation quality was...	fluctuating - consistent	2.93	2.25	3.18
During the term, translation improved...	not at all - clearly	3.29	3.00	3.36
The translation of general terms was...	bad - good	3.63	3.50	3.67
The translation of technical terms was...	bad - good	3.36	3.00	3.42
Usefulness		2.83	2.08	3.14
The translation helped me improve my performance in studying for this subject.	disagree - agree	2.71	2.00	3.00
The translation made it easier to follow the lecture.	disagree - agree	2.77	2.00	3.11
The translation made it easier to comprehend the content of the lecture.	disagree - agree	3.00	2.25	3.33

Table 31: Evaluation MT Component, **NG** non-German, **G** German

details. The interviews were subdivided into several parts, resembling those of the questionnaire.

Questions The first part asked about general information about the interviewee, including his/her mother tongue and language knowledge. The second part asked general questions about the system, including its usefulness, user behaviour, and whether it made it easier to follow the lecture. KIT also wanted to know in which situations it helped most. The third/forth part were about the STT component/the MT component, especially about obvious mistakes and usefulness. Finally, KIT also wanted to learn more about useful features, ideas, suggestions or specific problems.

Procedure Two foreign students, a female business-studies student from China and a male computer-science student from Ecuador accepted being interviewed after lectures during which

A translation service for the slides or the script.
A log or an archive of lectures to be downloaded.
Improvement of algorithms and quality.
Concentration on Chinese, as the Chinese are supposed to be the largest group of foreign students.

Table 32: Most frequent suggestions from students

Reduce the time lag of the LT and provide a more stable speed.
Reduce inaccurate translations, especially of technical terms.
Reduce difficulties with abbreviations and technical terms.

Table 33: Most frequently suggested improvements

A log or archive of lectures.
Automatic highlighting of some (rare) untranslated words, so that they can be faster recognised.
A possibility to add comments to the transcript and the translation.
The possibility to give direct feedback.
A greater number of languages available.

Table 34: Most frequently suggested improvements during interviews

they had used the lecture translator. The interview was recorded and transcribed.

Results The Chinese girl had a very good knowledge of German and no problems following lectures. She thus considered the LT little useful for herself. However, she thought that it was useful for foreign students with a lower level of German. She found that the speech recognition worked very well, the resulting transcript being of great use for students. Nevertheless, it was difficult for her to follow the slides, the lecturer, and the LT simultaneously, especially when there were charts, or when the lecturer was pointing at or explaining a picture. The other interviewee was of the same opinion. He considered the time lag too large. As it is even greater for the translation, the latter was subsequently considered less useful. Both mentioned mistakes in transcripts and translations, but they seemed being of minor importance to them. When asked for improvements, they expressed a multitude of ideas, including the ones shown in Table 34. The female respondent mentioned technical problems during one lecture, but explained that those were due to the unstable Internet connection in the lecture hall. Another idea was to add subtitles to videos that are recorded and published in some lectures, also outside the main lecture hall, and thus reaching more possible users. Concerning an archive, the two interviewees were not sure whether they preferred an off-line archive to the live function. Although an off-line archive would probably be useful for later studying, the live function as it is implemented today allowes to get more situational context information. Moreover, the Chinese female student told us that a lot of friends of hers do not actually go to lectures but study from home.

Thanks to this interview with two foreign students, KIT was able to gain further insight into their way of studying. KIT learnt that there are a lot of foreign students who do not actually go to lectures, but study from home. The students also provided information about the usefulness of the Lecture Translator and its components and expressed valuable ideas.

4.5.9 Individual feedback

During the terms, KIT got further feedback via email: One student expressed his thankfulness about the system. He especially appreciated the STT function and said he was interested in the future development of the system. Another student suggested to implement a countdown indicating the upcoming lectures, something KIT was able to rapidly integrate. He also asked for an issue list in case a lecture cannot be streamed, and an archive for passed lectures. A second student gave some feedback concerning the automatic correction of repetitions made by the lecturer himself. He also suggested some sort of status update, so that information about a lecture, e.g. "started 15 minutes ago" or "currently not available" can be displayed.

4.5.10 General results

All in all, the different methods allowed KIT to identify the strengths of the current system, some difficulties, and above all a lot of suggestions and ideas. Although feedback was too rare to be statistically relevant, some interesting aspects became obvious.

Feedback from the people that have used the system was mostly positive. They are definitely interested and thankful for the service. Still, some features can and will be implemented to further improve the user experience.

As to the questions KIT had asked in the beginning, namely "Does the tool help students?" "Is it easier to follow the lectures?" one can say that—at least based on the feedback and answers KIT got—that the tool does help (foreign) students. In its current form, however, it is sometimes difficult to switch between the lecturer, the presentation, and the LT, and still keep track of the situation. Thus, especially when a lecturer explains charts or pictures or takes notes, it might be a bit confusing. The integration of slides, an archive of the lectures and/or the possibility to annotate or highlight text would be of great help in this respect.

4.6 Voting session evaluation

4.6.1 RWTH

RWTH developed an automatic speech recognition system for the voting session evaluation based on two subsystems with short-term features which are augmented by deep multilingual bottleneck features. The final recognition result is obtained by system combination of the subsystems using confusion network decoding.

Subsystems were trained, based on the Mel-Frequency Cepstral Coefficients (MFCC) using a bank of 20 filters. For each time frame 16 coefficients including energy were extracted and a cepstral mean and variance normalization was applied on the segment level. Augmenting the MFCC features by a voicedness feature and applying a sliding window of size 9, 154-dimensional feature vectors were obtained that were projected down to 45 components using an LDA transformation. To introduce variability between the resulting subsystems and thereby improve the final system combination step, Perceptual Linear Predictive (PLP) features were also extracted in an analogous manner. Then phone-posterior-based features, estimated using a multilingual multilayer perceptron (MLP), were appended. The next subsection gives more details about those features which are part of the tandem approach where conventional short-term features and MLP features are combined in a single system.

The RWTH multilingual BN features are deep, hierarchical NNs trained on context-dependent triphone targets. Multiresolutional RASTA filter outputs are processed in a hierarchical way. Each hierarchy consists of 7-hidden-layer BN-MLPs and was trained on 1500 tied-triphone state targets per language. Since the BN is placed before the last hidden layer, RWTH also investigated the effect of introducing language dependent hidden layers after the bottleneck. The relative improvement compared to the target language BN features exceeds 5%. We updated the resulting language independent BN based English data only.

For both of our subsystems, MFCC and PLP features were normalized using Vocal Tract Length Normalization (VTLN). The VTLN warping factors were obtained by performing a grid search on the audio training data, then training a Gaussian classifier on the results. Eventually the classifier was applied to the training and recognition data to obtain the VTLNwarped features (fast-VTLN). To compensate for speaker variation, the Constrained Maximum Likelihood Linear Regression (CMLLR) technique was used in training and recognition. The adaptation matrices were estimated based on alignments computed using single Gaussians which in general

gives better results than full mixture models. For CMLLR, a two-pass recognition setup is necessary. Speakerlabels as required for CMLLR estimation were computed by clustering speech segments optimizing the Bayesian information criterion.

Based on the available training data, 4-gram language models (LMs) were estimated for each language using, smoothed by the Modified Kneser-Ney method. We partitioned the LM data into blocks, estimating n-gram probabilities for each block individually. Then the LMs were linearly interpolated while optimizing the perplexity on a holdout data set. Due to the changed domain of the recognition task the language model of the baseline system was adapted to the voting session domain by incorporating the text data extracted from voting sessions and using the TC-STAR languagemodel for the interpolated language model.

The acoustic models are based on triphones with cross-word context, modeled by a 3-state left-to-right hidden Markov model (HMM). A decision tree based state tying is applied resulting in a total of 4,500 generalized triphone states. The acoustic models consist of Gaussian mixture distributions with a globally pooled diagonal covariance matrix. Both, maximum likelihood (ML) and discriminative training are applied. For all acoustic models, the multilingual bottleneck features have been used within the tandem approach. In order to compensate for speaker variations we have used constrained maximum likelihood linear regression speaker adaptive training (SAT/CMLLR). The existing acoustic model was trained in in-house training data solely.

The recognition system operates with a time synchronous, word conditioned tree search, crossword decoder with batch cached likelihood calculations and efficient acoustic and language model lookaheads. A multi-pass recognition is built starting with a segmentation. In the first pass, a fast VTLN recognition is performed. The second pass consists of a CMLLR matrices adaptation to the first pass output followed by a recognition with SAT-CMLLR and MLLR matrices second pass output. In the final pass a system combination obtains the recognition result by confusion network decoding on the lattices resulting second pass recognition.

4.6.2 KIT

KIT based its system on the best English systems for the IWSLT2013. The systems feature 8000 context-dependent phones. We built two branches of systems using different phonessets. In each of these branches, we used different pre-processing setups. We combined Mel-frequency cepstral coefficients (MFCC) and minimum variance distortion response (MVDR) features with fundamental frequency variation (FFV) and Pitch as well as logMel with FFV and Pitch. Both pipelines output a 54 dimensional feature vector. Adjacent features are being stacked with a context of 6 and being fed into a neural network for the extraction of bottleneck features. This way, the network reduces the dimensionality from 702 down to 42. These features are then again stacked with a context of 6 and then an LDA is performed in order to reduce the dimensionality back to 42.

We adapted the original IWSLT systems for the new use case. In order to do so, we adapted the acoustic model in a semi-supervised fashion using 110 hours of voting session data. The data was automatically transcribed and aligned to the those transcripts. We interpolated the base language model from the IWSLT2013 evaluation with data from voting sessions, using a weight of 30%.

The final system was built training systems in several iterations. During the first pass, we combined the output of the systems of each branch using a CNC. We then adapted all systems on that output and performed a second CNC. In order to produce the final result, we combined the output of the two CNCs from the BEEP and CMU branch as well as the output from the best individual systems using ROVER.

In addition to the training and adaption of the acoustic model, we adapted the language model as well. We interpolated the existing language model with transcripts from voting sessions, using a weight of 0.3. By doing so, we could improve the WER on the performance on the 2013 dev set from initially 30.6% (unadapted) to 14.2%.

4.6.3 FBK

The acoustic models were based on a phone set for American English. Two sets of state-tied, cross-word, gender-independent triphone HMMs were trained for the first and second decoding step, respectively. Both models sets were speaker adaptively trained as described in Giuliani et al. (2006), Stemmer et al. (2005). AMs were trained on more than 300 hours of transcribed speech data from public corpora: about 143h of transcribed audio recordings from the NIST hub4 1996/1997 training data and the rest were from the EPPS collection (made available by the EU project TC-STAR) consisting of speeches delivered in English by politicians during the European Parliamentary Plenary Sessions.

The language model was trained on English data coming from different sources: newspapers, European Parliament, transcriptions of Voting sessions. In total about 2 Giga words were used to train a 4-grams LM having about 260 millions of 4-grams. The LM was pruned to build a manageable FSN, retaining about 40 millions of 4-grams. The lexicon was fixed to the 200K most frequent words.

4.6.4 Results

The results from the evaluation are shown in table 35. The numbers shown represent the WER on the stated data set.

Dataset	KIT	FBK	RWTH
dev2013	14.2%	18.8%	14.6%
eval2013	13.6%	16.5%	13.3%

Table 35: Results from the voting session evaluation

Bibliography

Auli, M., Galley, M., Quirk, C., and Zweig, G. (2013). Joint language and translation modeling with recurrent neural networks. In *Conference on Empirical Methods in Natural Language Processing*, pages 1044–1054, Seattle, WA, USA.

Axelrod, A., He, X., and Gao, J. (2011). Domain adaptation via pseudo in-domain data selection. In *Conference on Empirical Methods in Natural Language Processing*, pages 355–362, Edinburgh, United Kingdom.

Babaali, B., Serizel, R., Jalalvand, S., Falavigna, D., Gretter, R., and Giuliani, D. (2014). FBK @ IWSLT 2014 - ASR track. In *Proceedings of the International Workshop on Spoken Language Translation (IWSLT)*, pages 18–25.

Bell, P., Swietojanski, P., Driesen, J., Sinclair, M., McInnes, F., and Renals, S. (2014). The UEDIN ASR systems for the IWSLT 2014 evaluation. In *Proceedings of the International Workshop on Spoken Language Translation (IWSLT)*, pages 26–33.

Beloucif, M., Lo-kiu, C., and Wu, D. (2014). Improving MEANT based semantically tuned SMT. In *Proceedings of the International Workshop on Spoken Language Translation (IWSLT)*, pages 34–41.

Bertoldi, N., Mathur, P., Ruiz, N., and Federico, M. (2014). FBK's machine translation and speech translation systems for the IWSLT 2014 evaluation campaign. In *Proceedings of the International Workshop on Spoken Language Translation (IWSLT)*, pages 42–48.

Birch, A., Huck, M., Durrani, N., Bogoychev, N., and Koehn, P. (2014). Edinburgh SLT and MT system description for the IWSLT 2014 evaluation. In *Proceedings of the International Workshop on Spoken Language Translation (IWSLT)*, pages 49–56.

Bisazza, A., Ruiz, N., and Federico, M. (2011). Fill-up versus interpolation methods for phrase-based smt adaptation. In *International Workshop on Spoken Language Translation*, pages 136–143, San Francisco, CA, USA.

Cettolo, M., Girardi, C., and Federico, M. (2012). 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.

Cho, E., Niehues, J., and Waibel, A. (2012). Segmentation and punctuation prediction in speech language translation using a monolingual translation system. In *International Workshop on Spoken Language Translation*, pages 252–259, Hong Kong, China.

Clark, J., Dyer, C., Lavie, A., and Smith, N. (2011). Better hypothesis testing for statistical machine translation: Controlling for optimizer instability. In *Association for Computational Lingustics (ACL)*, Portland, OR, USA.

Do, Q. K., Herrmann, T., Niehues, J., Allauzen, A., Yvon, F., and Waibel, A. (2014). The kit-limsi translation system for wmt 2014. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 84–89, Baltimore, Maryland, USA. Association for Computational Linguistics.

Durrani, N., Fraser, A., and Schmid, H. (2013). Model with minimal translation units, but decode with phrases. In *Conference of the North American Chapter of the Association for Computational Linguistics*, Atlanta, GA, USA.

Durrani, N., Haddow, B., Koehn, P., and Heafield, K. (2014). Edinburghs phrase-based machine translation systems for wmt-14. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 97–104, Baltimore, Maryland, USA. Association for Computational Linguistics.

Durrani, N., Schmid, H., and Fraser, A. (2011). A joint sequence translation model with integrated reordering. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 1045–1054, Portland, Oregon, USA. Association for Computational Linguistics.

Freitag, M., Peitz, S., Wuebker, J., Ney, H., Huck, M., Sennrich, R., Durrani, N., Nadejde, M., Williams, P., Koehn, P., Herrmann, T., Cho, E., and Waibel, A. (2014a). Eu-bridge mt: Combined machine translation. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 105–113, Baltimore, Maryland, USA. Association for Computational Linguistics.

Freitag, M., Wübker, J., Peitz, S., Ney, H., Huck, M., Birch, A., Durrani, N., Koehn, P., Mediani, M., Slawik, I., Niehues, J., Cho, E., Waibel, A., Bertoldi, N., Cettolo, M., and Federico, M. (2014b). Combined spoken language translation. In *Proceedings of the International Workshop on Spoken Language Translation (IWSLT)*, pages 57–64.

Giuliani, D., Gerosa, M., and Brugnara, F. (2006). Improved automatic speech recognition through speaker normalization. *Computer, Speech and Language*, 20(1):107–123.

Graliński, F., Jassem, K., and Junczys-Dowmunt, M. (2013). Psi-toolkit: A natural language processing pipeline. In *Computational Linguistics*, pages 27–39. Springer.

Gretter, R. (2014a). Euronews: a multilingual benchmark for asr and lid. In *Proceedings of Interspeech*, Singapore.

Gretter, R. (2014b). Euronews: a multilingual speech corpus for asr. In *Proceedings of LREC*, Reykjavik, Iceland.

Heafield, K. (2011a). Kenlm: Faster and smaller language model queries. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 187–197, Edinburgh, Scotland. Association for Computational Linguistics.

Heafield, K. (2011b). KenLM: faster and smaller language model queries. In *EMNLP 2011 Sixth Workshop on Statistical Machine Translation*, pages 187–197, Edinburgh, Scotland, United Kingdom.

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

Herrmann, T., Mediani, M., Cho, E., Ha, T.-L., Niehues, J., Slawik, I., Zhang, Y., and Waibel, A. (2014). The karlsruhe institute of technology translation systems for the wmt 2014. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 130–135, Baltimore, Maryland, USA. Association for Computational Linguistics.

Kilgour, K., Heck, M., Müller, M., Sperber, M., Stücker, S., and Waibel, A. (2014). The 2014 kit IWSLT speech-to-text systems for English, German and Italian. In *Proceedings of the International Workshop on Spoken Language Translation (IWSLT)*, pages 73–79.

Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., Dyer, C. J., Bojar, O., Constantin, A., and Herbst, E. (2007). Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, pages 177–180, Prague, Czech Republic. Association for Computational Linguistics.

Lamel, L., Gauvain, J., and Adda, G. (2002). Lightly supervised and unsupervised acoustic model training. *Computer Speech and Language*, 16(1):115–129.

Moore, R. C. and Lewis, W. (2010). Intelligent selection of language model training data. In *Annual Meeting of the Association for Computational Linguistics*, pages 220–224, Uppsala, Sweden.

Niehues, J. and Waibel, A. (2012). Continuous space language models using restricted boltzmann machines. In *International Workshop on Spoken Language Translation*, Hong Kong, China.

Peitz, S., Freitag, M., Mauser, A., and Ney, H. (2011). Modeling punctuation prediction as machine translation. In *International Workshop on Spoken Language Translation*, pages 238–245, San Francisco, CA, USA.

Peitz, S., Wuebker, J., Freitag, M., and Ney, H. (2014). The rwth aachen german-english machine translation system for wmt 2014. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 157–162, Baltimore, Maryland, USA. Association for Computational Linguistics.

Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., Hannemann, M., Motlicek, P., Qian, Y., Schwarz, P., et al. (2011). The kaldi speech recognition toolkit. In *Proc. ASRU*, pages 1–4.

Schwenk, H., Rousseau, A., and Attik, M. (2012). Large, Pruned or Continuous Space Language Models on a GPU for Statistical Machine Translation. In *NAACL-HLT 2012 Workshop: Will We Ever Really Replace the N-gram Model? On the Future of Language Modeling for HLT*, pages 11–19, Montréal, Canada.

Slawik, I., Mediani, M., Niehues, J., Zhang, Y., Cho, E., Herrmann, T., Ha, T.-L., and Waibel, A. (2014). The KIT translation systems for IWSLT 2014. In *Proceedings of the International Workshop on Spoken Language Translation (IWSLT)*, pages 119–126.

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

Stolke, A. (2002). SRILM - an extensible language modeling toolkit. In *Proceedings of the International Conference on Spoken Language Processing*.

Sundermeyer, M., Alkhouli, T., Wuebker, J., and Ney, H. (2014). Translation modeling with bidirectional recurrent neural networks. In *Conference on Empirical Methods in Natural Language Processing*, pages 14–25, Doha, Qatar.

Williams, P., Sennrich, R., Nadejde, M., Huck, M., Hasler, E., and Koehn, P. (2014). Edinburghs syntax-based systems at wmt 2014. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 207–214, Baltimore, Maryland, USA. Association for Computational Linguistics.

Wolk, K. and Marasek, K. (2013). Polish-English speech statistical machine translation systems for the IWSLT 2013. In *Proceedings of the International Workshop on Spoken Language Translation (IWSLT)*.

Wolk, K. and Marasek, K. (2014). Polish - English speech statistical machine translation systems for the IWSLT 2014. In *Proceedings of the International Workshop on Spoken Language Translation (IWSLT)*, pages 143–149.

Wołk, K. and Marasek, K. (2014). Real-time statistical speech translation. In *New Perspectives in Information Systems and Technologies, Volume 1*, pages 107–113. Springer.

Wołk, K. and Marasek, K. (2015). Polish-english statistical machine translation of medical texts. In *New Research in Multimedia and Internet Systems*, pages 169–179. Springer.

Wübker, J., Peitz, S., Guta, A., and Ney, H. (2014). The RWTH Aachen machine translation systems for IWSLT 2014. In *Proceedings of the International Workshop on Spoken Language Translation (IWSLT)*, pages 150–155.

Wuebker, J., Peitz, S., Rietig, F., and Ney, H. (2013). Improving statistical machine translation with word class models. In *Conference on Empirical Methods in Natural Language Processing*, pages 1377–1381, Seattle, WA, USA.

Appendix A Details of significance tests

A.1 Output of sc_stats on the Euronews ASR evaluation

A.1.1 Arabic

```
,----------------------------------------------------------.
```

```
| Composite Report of All Significance Tests                        |
| For the  Test                                                     |
|                                                                   |
| Test Name                                                Abbrev.|
|-----------------------------------------------------------  -----|
| Matched Pair Sentence Segment (Word Error)                  MP  |
| Signed Paired Comparison (Speaker Word Error Rate (%))      SI  |
| Wilcoxon Signed Rank (Speaker Word Error Rate (%))          WI  |
| McNemar (Sentence Error)                                    MN  |
|                                                                 |
|                                                                 |
|-----------------------------------------------------------------|
|  Test   ||      | PEV |    FBK       |  RWTH                     |
| Abbrev. ||      |     |              |                           |
|---------++----------------------+---------------------------|
|   MP    || PEV  |     | ~    0.250   | ~    0.834                |
|   SI    ||      |     | ~    0.581   | ~    1.000                |
|   WI    ||      |     | ~    0.134   | ~    0.555                |
|   MN    ||      |     | PEV   <0.001 | PEV  <0.001               |
|---------++----------------------+---------------------------|
|   MP    || FBK  |     |              | ~    0.197                |
|   SI    ||      |     |              | ~    0.581                |
|   WI    ||      |     |              | ~    0.810                |
|   MN    ||      |     |              | RWTH   <0.001             |
|-----------------------------------------------------------------|
| These significance tests are all two-tailed tests with the      |
| null hypothesis that there is no performance difference         |
| between the wo systems.                                         |
|                                                                 |
| The first column indicates if the test finds a significant      |
| difference at the level of p=0.05.  It consists of '~' if no    |
| difference is found at this significance level. If a            |
| difference at this level is found, this column indicates the    |
| system with the higher value on the performance statistic       |
| utilized by the particular test.                                |
|                                                                 |
| The second column specifies the minimum value of p for which    |
| the test finds a significant difference at the level of p.      |
|                                                                 |
| The third column indicates if the test finds a significant      |
| difference at the level of p=0.001 (""), at the level of        |
| p=0.01, but not p=0.001 (""), or at the level of p=0.05, but    |
| not p=0.01 ("").                                                |
|                                                                 |
| A test finds significance at level p if, assuming the null      |
| hypothesis, the probability of the test statistic having a      |
| value at least as extreme as that actually found, is no more    |
| than p.                                                         |
`-----------------------------------------------------------------'
```

A.1.2 English

```
,-----------------------------------------------------------------.
```

```
| Composite Report of All Significance Tests                    |
| For the  Test                                                 |
|                                                               |
| Test Name                                              Abbrev.|
|-------------------------------------------------------  -----|
| Matched Pair Sentence Segment (Word Error)               MP   |
| Signed Paired Comparison (Speaker Word Error Rate (%))   SI   |
| Wilcoxon Signed Rank (Speaker Word Error Rate (%))       WI   |
| McNemar (Sentence Error)                                 MN   |
|                                                               |
|                                                               |
|---------------------------------------------------------------|
|  Test   ||      |  FBK          |        PEV                   |
| Abbrev. ||      |               |                              |
|---------++------+---------------+------------------+-------|
|   MP    || KIT  | ~    0.472    | KIT    <0.001                |
|   SI    ||      | FBK    <0.001 | KIT    <0.001                |
|   WI    ||      | FBK     0.032 | KIT     0.022                |
|   MN    ||      | FBK    <0.001 | KIT    <0.001                |
|---------++------+---------------+------------------+-------|
|   MP    || FBK  |               | FBK    <0.001                |
|   SI    ||      |               | FBK    <0.001                |
|   WI    ||      |               | FBK    <0.001                |
|   MN    ||      |               | FBK    <0.001                |
|---------------------------------------------------------------|
| These significance tests are all two-tailed tests with the    |
| null hypothesis that there is no performance difference       |
| between the wo systems.                                       |
|                                                               |
| The first column indicates if the test finds a significant    |
| difference at the level of p=0.05.  It consists of '~' if no  |
| difference is found at this significance level. If a          |
| difference at this level is found, this column indicates the  |
| system with the higher value on the performance statistic     |
| utilized by the particular test.                              |
|                                                               |
| The second column specifies the minimum value of p for which  |
| the test finds a significant difference at the level of p.    |
|                                                               |
| The third column indicates if the test finds a significant    |
| difference at the level of p=0.001 (""), at the level of      |
| p=0.01, but not p=0.001 (""), or at the level of p=0.05, but  |
| not p=0.01 ("").                                              |
|                                                               |
| A test finds significance at level p if, assuming the null    |
| hypothesis, the probability of the test statistic having a    |
| value at least as extreme as that actually found, is no more  |
| than p.                                                       |
`---------------------------------------------------------------'
```

A.1.3 Italian

```
,---------------------------------------------------------------.
```

```
| Composite Report of All Significance Tests                    |
| For the  Test                                                 |
|                                                               |
| Test Name                                              Abbrev.|
|-------------------------------------------------------  -----|
| Matched Pair Sentence Segment (Word Error)               MP   |
| Signed Paired Comparison (Speaker Word Error Rate (%))   SI   |
| Wilcoxon Signed Rank (Speaker Word Error Rate (%))       WI   |
| McNemar (Sentence Error)                                 MN   |
|                                                               |
|                                                               |
|---------------------------------------------------------------|
|  Test   ||     | FBK        | PEV                           |
| Abbrev. ||     |            |                               |
|---------++-----------------+--------------------+-------|
|   MP    || KIT | FBK <0.001 | PEV <0.001                    |
|   SI    ||     | FBK <0.001 | PEV <0.001                    |
|   WI    ||     | FBK <0.001 | PEV <0.001                    |
|   MN    ||     | FBK <0.001 | PEV <0.001                    |
|---------++-----------------+--------------------+-------|
|   MP    || FBK |            | ~    0.873                    |
|   SI    ||     |            | PEV <0.001                    |
|   WI    ||     |            | ~    0.912                    |
|   MN    ||     |            | PEV <0.001                    |
|---------------------------------------------------------------|
| These significance tests are all two-tailed tests with the    |
| null hypothesis that there is no performance difference       |
| between the wo systems.                                       |
|                                                               |
| The first column indicates if the test finds a significant    |
| difference at the level of p=0.05.  It consists of '~' if no  |
| difference is found at this significance level. If a          |
| difference at this level is found, this column indicates the  |
| system with the higher value on the performance statistic     |
| utilized by the particular test.                              |
|                                                               |
| The second column specifies the minimum value of p for which  |
| the test finds a significant difference at the level of p.    |
|                                                               |
| The third column indicates if the test finds a significant    |
| difference at the level of p=0.001 (""), at the level of      |
| p=0.01, but not p=0.001 (""), or at the level of p=0.05, but  |
| not p=0.01 ("").                                              |
|                                                               |
| A test finds significance at level p if, assuming the null    |
| hypothesis, the probability of the test statistic having a    |
| value at least as extreme as that actually found, is no more  |
| than p.                                                       |
`---------------------------------------------------------------'
```

A.1.4 Polish

```
,---------------------------------------------------------------.
```

```
| Composite Report of All Significance Tests                    |
| For the  Test                                                 |
|                                                               |
| Test Name                                              Abbrev.|
|-------------------------------------------------------  -----|
| Matched Pair Sentence Segment (Word Error)               MP   |
| Signed Paired Comparison (Speaker Word Error Rate (%))   SI   |
| Wilcoxon Signed Rank (Speaker Word Error Rate (%))       WI   |
| McNemar (Sentence Error)                                 MN   |
|                                                               |
|                                                               |
|---------------------------------------------------------------|
|  Test   || RWTH                                               |
| Abbrev. ||                                                    |
|---------++----------------------+----------------------+---|
|   MP    || RWTH <0.001                                        |
|   SI    || ~     0.219                                        |
|   WI    || RWTH  0.047                                        |
|   MN    || RWTH <0.001                                        |
|---------------------------------------------------------------|
| These significance tests are all two-tailed tests with the    |
| null hypothesis that there is no performance difference       |
| between the wo systems.                                       |
|                                                               |
| The first column indicates if the test finds a significant    |
| difference at the level of p=0.05.  It consists of '~' if no  |
| difference is found at this significance level. If a          |
| difference at this level is found, this column indicates the  |
| system with the higher value on the performance statistic     |
| utilized by the particular test.                              |
|                                                               |
| The second column specifies the minimum value of p for which  |
| the test finds a significant difference at the level of p.    |
|                                                               |
| The third column indicates if the test finds a significant    |
| difference at the level of p=0.001 (""), at the level of      |
| p=0.01, but not p=0.001 (""), or at the level of p=0.05, but  |
| not p=0.01 ("").                                              |
|                                                               |
| A test finds significance at level p if, assuming the null    |
| hypothesis, the probability of the test statistic having a    |
| value at least as extreme as that actually found, is no more  |
| than p.                                                       |
'---------------------------------------------------------------'
```

A.1.5 Portuguese

```
,---------------------------------------------------------------.
| Composite Report of All Significance Tests                    |
| For the  Test                                                 |
|                                                               |
| Test Name                                              Abbrev.|
|-------------------------------------------------------  -----|
```

```
| Matched Pair Sentence Segment (Word Error)                MP  |
| Signed Paired Comparison (Speaker Word Error Rate (%))    SI  |
| Wilcoxon Signed Rank (Speaker Word Error Rate (%))        WI  |
| McNemar (Sentence Error)                                  MN  |
|                                                               |
|                                                               |
|  Test   ||       | PEV          |        RWTH                 |
| Abbrev. ||       |                                            |
|---------++-------+--------------+---------------------+------|
|   MP    || KIT   | KIT  <0.001  | ~     0.653                 |
|   SI    ||       | KIT  <0.001  | KIT  <0.001                 |
|   WI    ||       | KIT  <0.001  | ~     0.757                 |
|   MN    ||       | KIT  <0.001  | RWTH <0.001                 |
|---------++-------+--------------+---------------------+------|
|   MP    || PEV   |              | RWTH <0.001                 |
|   SI    ||       |              | RWTH <0.001                 |
|   WI    ||       |              | RWTH <0.001                 |
|   MN    ||       |              | RWTH <0.001                 |
|---------------------------------------------------------------|
| These significance tests are all two-tailed tests with the    |
| null hypothesis that there is no performance difference       |
| between the wo systems.                                       |
|                                                               |
| The first column indicates if the test finds a significant    |
| difference at the level of p=0.05.  It consists of '~' if no  |
| difference is found at this significance level. If a          |
| difference at this level is found, this column indicates the  |
| system with the higher value on the performance statistic     |
| utilized by the particular test.                              |
|                                                               |
| The second column specifies the minimum value of p for which  |
| the test finds a significant difference at the level of p.    |
|                                                               |
| The third column indicates if the test finds a significant    |
| difference at the level of p=0.001 (""), at the level of      |
| p=0.01, but not p=0.001 (""), or at the level of p=0.05, but  |
| not p=0.01 ("").                                              |
|                                                               |
| A test finds significance at level p if, assuming the null    |
| hypothesis, the probability of the test statistic having a    |
| value at least as extreme as that actually found, is no more  |
| than p.                                                       |
'---------------------------------------------------------------'
```

A.1.6 Russian

```
,---------------------------------------------------------------.
| Composite Report of All Significance Tests                    |
| For the  Test                                                 |
|                                                               |
| Test Name                                              Abbrev.|
|-------------------------------------------------------  -----|
| Matched Pair Sentence Segment (Word Error)                MP  |
```

```
| Signed Paired Comparison (Speaker Word Error Rate (%))   SI  |
| Wilcoxon Signed Rank (Speaker Word Error Rate (%))       WI  |
| McNemar (Sentence Error)                                 MN  |
|                                                              |
|                                                              |

|-------------------------------------------------------------|
|  Test   ||                    | FBK                        |
| Abbrev. ||                    |                             |
|---------++--------------------+---------------------+-------|
|   MP    || KIT                | KIT <0.001                  |
|   SI    ||                    | KIT <0.001                  |
|   WI    ||                    | KIT <0.001                  |
|   MN    ||                    | KIT <0.001                  |
|---------++--------------------+---------------------+-------|
| These significance tests are all two-tailed tests with the   |
| null hypothesis that there is no performance difference      |
| between the wo systems.                                      |
|                                                              |
| The first column indicates if the test finds a significant   |
| difference at the level of p=0.05.  It consists of '~' if no |
| difference is found at this significance level. If a         |
| difference at this level is found, this column indicates the |
| system with the higher value on the performance statistic    |
| utilized by the particular test.                             |
|                                                              |
| The second column specifies the minimum value of p for which |
| the test finds a significant difference at the level of p.   |
|                                                              |
| The third column indicates if the test finds a significant   |
| difference at the level of p=0.001 (""), at the level of     |
| p=0.01, but not p=0.001 (""), or at the level of p=0.05, but |
| not p=0.01 ("").                                             |
|                                                              |
| A test finds significance at level p if, assuming the null   |
| hypothesis, the probability of the test statistic having a   |
| value at least as extreme as that actually found, is no more |
| than p.                                                      |
'-------------------------------------------------------------'
```

A.1.7 Turkish

```
,-------------------------------------------------------------.
| Composite Report of All Significance Tests                   |
| For the  Test                                                |
|                                                              |
| Test Name                                            Abbrev.|
|-----------------------------------------------------  -----|
| Matched Pair Sentence Segment (Word Error)             MP    |
| Signed Paired Comparison (Speaker Word Error Rate (%))   SI  |
| Wilcoxon Signed Rank (Speaker Word Error Rate (%))       WI  |
| McNemar (Sentence Error)                                 MN  |
|                                                              |
```

```
|                                                                  |
|---------++--------------------+--------------------+-------|
|  Test   ||     | FBK                                          |
| Abbrev. ||     |                                              |
|---------++--------------------+--------------------+-------|
|   MP    || KIT | ~    0.165                                   |
|   SI    ||     | FBK <0.001                                   |
|   WI    ||     | ~    0.535                                   |
|   MN    ||     | FBK <0.001                                   |
|---------++--------------------+--------------------+-------|
| These significance tests are all two-tailed tests with the   |
| null hypothesis that there is no performance difference      |
| between the wo systems.                                      |
|                                                              |
| The first column indicates if the test finds a significant   |
| difference at the level of p=0.05.  It consists of '~' if no |
| difference is found at this significance level. If a         |
| difference at this level is found, this column indicates the |
| system with the higher value on the performance statistic    |
| utilized by the particular test.                             |
|                                                              |
| The second column specifies the minimum value of p for which |
| the test finds a significant difference at the level of p.   |
|                                                              |
| The third column indicates if the test finds a significant   |
| difference at the level of p=0.001 (""), at the level of     |
| p=0.01, but not p=0.001 (""), or at the level of p=0.05, but |
| not p=0.01 ("").                                             |
|                                                              |
| A test finds significance at level p if, assuming the null   |
| hypothesis, the probability of the test statistic having a   |
| value at least as extreme as that actually found, is no more |
| than p.                                                      |
'--------------------------------------------------------------'
```

A.2 Output of sc_stats on the Skynews ASR evaluation

```
,-------------------------------------------------------------------------.
| Composite Report of All Significance Tests                              |
| For the  Test                                                           |
|                                                                         |
| Test Name                                                 Abbrev.       |
|-----------------------------------------------------------  -----       |
| Matched Pair Sentence Segment (Word Error)                  MP          |
| Signed Paired Comparison (Speaker Word Error Rate (%))      SI          |
| Wilcoxon Signed Rank (Speaker Word Error Rate (%))          WI          |
| McNemar (Sentence Error)                                    MN          |
|                                                                         |
|                                                                         |
-------------------------------------------------------------------------|
|  Test   ||      | kit       | pev         | rwth        | uedin         |
| Abbrev. ||      |           |             |             |               |
|---------++------------------+-------------+-------------+---------------|
```

```
|   MP    ||  fbk | ~   0.920 | fbk  <0.001 | ~    0.952  | fbk   <0.001  |
|   SI    ||      | ~   1.000 | fbk   0.016 | ~    0.453  | ~      0.125  |
|   WI    ||      | ~   0.873 | fbk   0.018 | ~    0.873  | fbk    0.043  |
|   MN    ||      | ~   1.000 | ~     1.000 | ~    1.000  | ~      1.000  |
|---------++------+-----------+-------------+-------------+---------------|
|   MP    ||  kit |           | kit <0.001  | ~    0.952  | kit   <0.001  |
|   SI    ||      |           | kit  0.016  | ~    1.000  | kit    0.016  |
|   WI    ||      |           | kit  0.018  | ~    1.000  | kit    0.018  |
|   MN    ||      |           | ~    1.000  | ~    1.000  | ~      1.000  |
|---------++------+-----------+-------------+-------------+---------------|
|   MP    ||  pev |           |             | rwth <0.001 | uedin <0.001  |
|   SI    ||      |           |             | rwth  0.016 | uedin  0.016  |
|   WI    ||      |           |             | rwth  0.018 | uedin  0.018  |
|   MN    ||      |           |             | ~     1.000 | ~      1.000  |
|---------++------+-----------+-------------+-------------+---------------|
|   MP    || rwth |           |             |             | rwth   <0.001 |
|   SI    ||      |           |             |             | ~    0.453    |
|   WI    ||      |           |             |             | ~    0.091    |
|   MN    ||      |           |             |             | ~    1.000    |
|---------++------+-----------+-------------+-------------+---------------|
| These significance tests are all two-tailed tests with the              |
| null hypothesis that there is no performance difference                 |
| between the wo systems.                                                 |
|                                                                         |
| The first column indicates if the test finds a significant              |
| difference at the level of p=0.05.  It consists of '~' if no            |
| difference is found at this significance level. If a                    |
| difference at this level is found, this column indicates the            |
| system with the higher value on the performance statistic               |
| utilized by the particular test.                                        |
|                                                                         |
| The second column specifies the minimum value of p for which            |
| the test finds a significant difference at the level of p.              |
|                                                                         |
| The third column indicates if the test finds a significant              |
| difference at the level of p=0.001 (""), at the level of                |
| p=0.01, but not p=0.001 (""), or at the level of p=0.05, but            |
| not p=0.01 ("").                                                        |
|                                                                         |
| A test finds significance at level p if, assuming the null              |
| hypothesis, the probability of the test statistic having a              |
| value at least as extreme as that actually found, is no more            |
| than p.                                                                 |
'-------------------------------------------------------------------------'
```

A.3 Output of sc_stats on the IWSLT ASR evaluation

A.3.1 English

```
,-----------------------------------------------------------------------------------------------------.
| Composite Report of All Significance Tests                                                          |
| For the  Test                                                                                       |
|                                                                                                     |
| Test Name                                        Abbrev.                                            |
|--------------------------------------------------  -----                                            |
| Matched Pair Sentence Segment (Word Error)          MP                                              |
```

Test Name	Abbrev.
Signed Paired Comparison (Speaker Word Error Rate (%))	SI
Wilcoxon Signed Rank (Speaker Word Error Rate (%))	WI
McNemar (Sentence Error)	MN

Test Abbrev.		ioit	kit	lium	mitll-afrl	nict	uedin
MP	fbk	fbk <0.001	~ 0.772	fbk 0.001	mitll-afrl <0.001	nict <0.001	fbk <0.001
SI		fbk <0.001	~ 0.607	~ 0.607	mitll-afrl <0.001	nict <0.001	fbk 0.035
WI		fbk <0.001	~ 0.826	~ 0.089	mitll-afrl 0.002	nict <0.001	fbk 0.032
MN		fbk <0.001	fbk <0.001	fbk <0.001	mitll-afrl <0.001	nict <0.001	fbk <0.001
MP	ioit		kit <0.001	lium <0.001	mitll-afrl <0.001	nict <0.001	uedin <0.001
SI			kit <0.001	lium <0.001	mitll-afrl <0.001	nict <0.001	uedin <0.001
WI			kit <0.001	lium <0.001	mitll-afrl <0.001	nict <0.001	uedin <0.001
MN			kit <0.001	lium <0.001	mitll-afrl <0.001	nict <0.001	uedin <0.001
MP	kit			kit <0.001	mitll-afrl <0.001	nict <0.001	kit <0.001
SI				~ 0.302	mitll-afrl 0.035	nict <0.001	kit 0.035
WI				~ 0.112	mitll-afrl 0.007	nict 0.003	kit 0.015
MN				kit <0.001	mitll-afrl <0.001	nict <0.001	kit <0.001
MP	lium				mitll-afrl <0.001	nict <0.001	~ 0.219
SI					mitll-afrl <0.001	nict <0.001	~ 0.607
WI					mitll-afrl 0.001	nict <0.001	~ 0.697
MN					mitll-afrl <0.001	nict <0.001	lium <0.001
MP	mitll-afrl					nict <0.001	mitll-afrl <0.001
SI						nict 0.007	mitll-afrl <0.001
WI						nict 0.004	mitll-afrl 0.001
MN						nict <0.001	mitll-afrl <0.001
MP	nict						nict <0.001
SI							nict <0.001
WI							nict <0.001
MN							nict <0.001

These significance tests are all two-tailed tests with the null hypothesis that there is no performance difference between the wo systems.

The first column indicates if the test finds a significant difference at the level of p=0.05. It consists of '~' if no difference is found at this significance level. If a difference at this level is found, this column indicates the system with the higher value on the performance statistic utilized by the particular test.

The second column specifies the minimum value of p for which the test finds a significant difference at the level of p.

The third column indicates if the test finds a significant difference at the level of p=0.001 (""), at the level of p=0.01, but not p=0.001 (""), or at the level of p=0.05, but not p=0.01 ("").

A test finds significance at level p if, assuming the null hypothesis, the probability of the test statistic having a value at least as extreme as that actually found, is no more than p.

A.3.2 German

Composite Report of All Significance Tests
For the Test

Test Name	Abbrev.
Matched Pair Sentence Segment (Word Error)	MP
Signed Paired Comparison (Speaker Word Error Rate (%))	SI
Wilcoxon Signed Rank (Speaker Word Error Rate (%))	WI
McNemar (Sentence Error)	MN

Test Abbrev.		kit	uedin

```
|   MP    || fbk | kit <0.001 | uedin <0.001                    |
|   SI    ||     | kit  0.002 | uedin  0.002                    |
|   WI    ||     | kit  0.005 | uedin  0.005                    |
|   MN    ||     | kit <0.001 | uedin <0.001                    |
|---------++-----+------------+---------------------------------|
|   MP    || kit |            | kit   <0.001                    |
|   SI    ||     |            | kit    0.002                    |
|   WI    ||     |            | kit    0.005                    |
|   MN    ||     |            | kit   <0.001                    |
|---------++-----+------------+---------------------------------|
| These significance tests are all two-tailed tests with the    |
| null hypothesis that there is no performance difference       |
| between the wo systems.                                       |
|                                                               |
| The first column indicates if the test finds a significant    |
| difference at the level of p=0.05.  It consists of '~' if no  |
| difference is found at this significance level. If a          |
| difference at this level is found, this column indicates the  |
| system with the higher value on the performance statistic     |
| utilized by the particular test.                              |
|                                                               |
| The second column specifies the minimum value of p for which  |
| the test finds a significant difference at the level of p.    |
|                                                               |
| The third column indicates if the test finds a significant    |
| difference at the level of p=0.001 (""), at the level of      |
| p=0.01, but not p=0.001 (""), or at the level of p=0.05, but  |
| not p=0.01 ("").                                              |
|                                                               |
| A test finds significance at level p if, assuming the null    |
| hypothesis, the probability of the test statistic having a    |
| value at least as extreme as that actually found, is no more  |
| than p.                                                       |
`---------------------------------------------------------------'
```

A.3.3 Italian

```
,-------------------------------------------------------------------------.
| Composite Report of All Significance Tests                              |
| For the  Test                                                           |
|                                                                         |
| Test Name                                              Abbrev.          |
|-------------------------------------------------------  ---------------- |
| Matched Pair Sentence Segment (Word Error)               MP             |
| Signed Paired Comparison (Speaker Word Error Rate (%))   SI             |
| Wilcoxon Signed Rank (Speaker Word Error Rate (%))       WI             |
| McNemar (Sentence Error)                                 MN             |
|                                                                         |
|                                                                         |
|-------------------------------------------------------------------------|
|  Test   ||           | kit         | mitll-afrl        |vecsys-lium      |
| Abbrev. ||           |             |                   |                 |
|---------++-----------------------------------------------+-----------------
```

```
|   MP    || fbk        | fbk <0.001  | ~           0.072 | vecsys-lium <0.001 |
|   SI    ||            | ~    0.581  | ~           0.581 | vecsys-lium  0.022 |
|   WI    ||            | ~    0.384  | ~           0.555 | vecsys-lium  0.016 |
|   MN    ||            | fbk <0.001  | mitll-afrl <0.001 | vecsys-lium <0.001 |
|---------++------------------------------------------------+------------------
|   MP    || kit        |             | mitll-afrl <0.001 | vecsys-lium <0.001 |
|   SI    ||            |             | ~           0.267 | ~            0.267 |
|   WI    ||            |             | ~           0.280 | ~            0.134 |
|   MN    ||            |             | mitll-afrl <0.001 | vecsys-lium <0.001 |
|---------++------------------------------------------------+------------------
|   MP    || mitll-afrl |             |                   | vecsys-lium  0.001 |
|   SI    ||            |             |                   | ~            0.267 |
|   WI    ||            |             |                   | ~            0.055 |
|   MN    ||            |             |                   | mitll-afrl  <0.001 |
|---------++-------------------------------------------------------------------|
| These significance tests are all two-tailed tests with the                    |
| null hypothesis that there is no performance difference                       |
| between the wo systems.                                                       |
|                                                                               |
| The first column indicates if the test finds a significant                    |
| difference at the level of p=0.05.  It consists of '~' if no                  |
| difference is found at this significance level. If a                          |
| difference at this level is found, this column indicates the                  |
| system with the higher value on the performance statistic                     |
| utilized by the particular test.                                              |
|                                                                               |
| The second column specifies the minimum value of p for which                  |
| the test finds a significant difference at the level of p.                    |
|                                                                               |
| The third column indicates if the test finds a significant                    |
| difference at the level of p=0.001 (""), at the level of                      |
| p=0.01, but not p=0.001 (""), or at the level of p=0.05, but                  |
| not p=0.01 ("").                                                              |
|                                                                               |
| A test finds significance at level p if, assuming the null                    |
| hypothesis, the probability of the test statistic having a                    |
| value at least as extreme as that actually found, is no more                  |
| than p.                                                                       |
'-------------------------------------------------------------------------------'
```

A.4 Output of the significance test calculations for the IWSLT MT and SLT evaluation

```
#### multeval-0.5.1

MT
        de-en
        2013                              2014
        BLEU      stdev    p-value      BLEU      stdev    p-value
EU-BR   29.919930 0.535844 -            26.231261 0.473442 -
RWTH    28.916094 0.536604 0.000100     25.516706 0.464712 0.000400
KIT     28.070049 0.527605 0.000100     24.512901 0.453105 0.000100
NAIST   28.645853 0.537212 0.000300     24.085817 0.444451 0.000100
```

UEDIN	28.864529	0.536241	0.000600	23.534784	0.447253	0.000100
FBK	26.393532	0.521663	0.000100	20.736752	0.423835	0.000100
KLE	24.092730	0.492717	0.000100	19.413836	0.420354	0.000100

en-de

	2013			2014		
	BLEU	stdev	p-value	BLEU	stdev	p-value
EU-BR	26.172242	0.602676	-	23.137378	0.500089	-
KIT	25.949673	0.592128	0.494051	22.561278	0.494394	0.033197
UEDIN	25.262253	0.599033	0.000800	22.534262	0.486010	0.003000
NAIST	25.696670	0.586607	0.181182	22.031048	0.474432	0.000100
KLE	21.645827	0.570914	0.000100	19.181598	0.453986	0.000100

en-fr

	2013			2014		
	BLEU	stdev	p-value	BLEU	stdev	p-value
EU-BR	41.824843	0.648497	-	38.352275	0.589494	-
KIT	41.422991	0.652288	0.286871	37.647114	0.593532	0.008499
UEDIN	40.938298	0.682627	0.008099	37.305396	0.584752	0.000800
RWTH	40.889037	0.648382	0.000600	37.118007	0.581380	0.000100
MITLL	40.396231	0.652105	0.000200	37.018608	0.583501	0.000100
FBK	39.425376	0.637622	0.000100	35.679985	0.561805	0.000100
MIRACL	30.671469	0.591806	0.000100	27.059732	0.491746	0.000100
SFAX				17.136986	0.399697	0.000100

SLT

de-en

	2014		
	BLEU	stdev	p-value
EU-BR	19.188170	0.405544	-
KIT	18.187792	0.395026	0.000100
UEDIN	17.864032	0.401957	0.000100
RWTH	17.564935	0.405188	0.000100
KLE	10.006034	0.317593	0.000100

en-de

	2014		
	BLEU	stdev	p-value
KIT	17.042712	0.454055	-
UEDIN	17.020886	0.465449	0.965503
USFD	14.720794	0.428437	0.000100
KLE	13.011356	0.397054	0.000100

en-fr

	2014		
	BLEU	stdev	p-value
KIT	28.865311	0.569651	-
LIUM	28.360707	0.571438	0.131387
RWTH	28.333045	0.578700	0.091091
FBK	27.065674	0.555414	0.000100
UEDIN	26.852143	0.561879	0.000100
LIMSI	26.699834	0.570063	0.000100

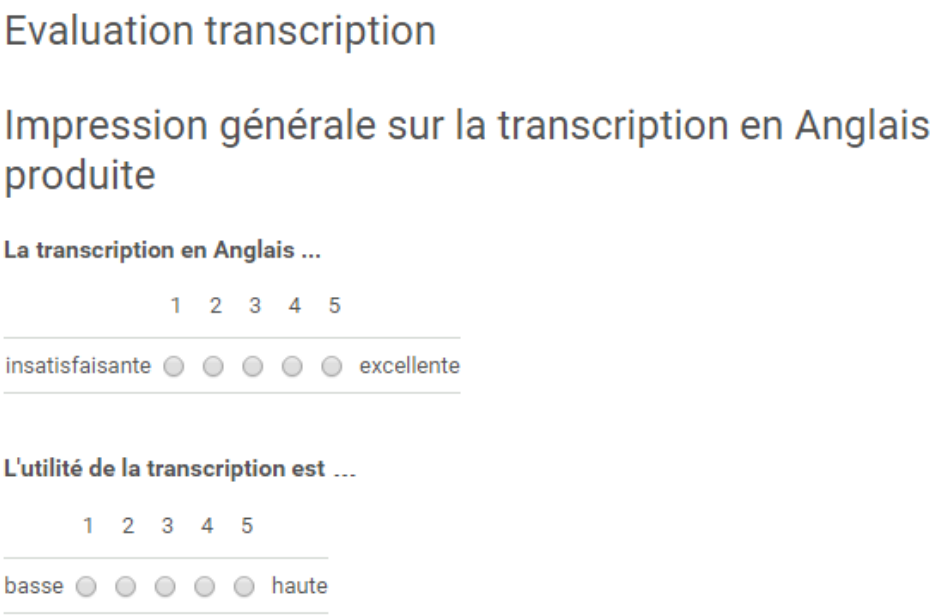
```
USFD     24.696983 0.530973 0.000100
```

A.5 Webinar feedback form

There are two types of forms used during the field test, the webinar specific feedback forms (webinar form) filled after each webinar viewing and the overall feedback forms (overall form) filled after all webinars have been viewed, which led to a summary feedback.

The feedback forms are in French as they have been filled by French speaking testers, listening to webinars in English, webinars that were automatically transcribed into English and translated into French.

A.5.1 Feedback form for transcription / translation (per webinar)



Qualité

Les erreurs de la transcription en Anglais sont ...

	1	2	3	4	5	
distrayantes	○	○	○	○	○	non distrayantes

Le délais de transcription est ...

	1	2	3	4	5	
grand	○	○	○	○	○	petit

la transcription est ...

	1	2	3	4	5	
hâchée	○	○	○	○	○	fluide

la transcription de termes généraux est ...

	1	2	3	4	5	
mauvaise	○	○	○	○	○	bonne

la transcription de termes techniques est ...

	1	2	3	4	5	
mauvaise	○	○	○	○	○	bonne

Utilité perçue de la transcription

La transcription me permet de mieux comprendre la matière présentée

	1	2	3	4	5	
désapprouve	○	○	○	○	○	approuve

La transcription me permet de mieux suivre la matière présentée

	1	2	3	4	5	
désapprouve	○	○	○	○	○	approuve

Question sur le contenu

A combien de questions-clés doit-on trouver une réponse pour savoir comment une compagnie se différencie de la compétition? *

Evaluation Traduction

Impression générale sur la traduction en Français produite

La traduction en Francais ...

	1	2	3	4	5	
insatisfaisante	○	○	○	○	○	excellente

L'utilité de la traduction est ...

	1	2	3	4	5	
distrayantes	○	○	○	○	○	non distrayantes

Qualité

Les erreurs de la traduction en Français sont ...

	1	2	3	4	5	
hâchée	○	○	○	○	○	fluide

Le délais de traduction est ...

	1	2	3	4	5	
grand	○	○	○	○	○	petit

la traduction est ...

	1	2	3	4	5	
mauvaise	○	○	○	○	○	bonne

la traduction de termes généraux est ...

	1	2	3	4	5	
mauvaise	○	○	○	○	○	bonne

la traduction de termes techniques est ...

	1	2	3	4	5	
basse	○	○	○	○	○	haute

Utilité perçue de la traduction

La traduction me permet de mieux comprendre la matière présentée

	1	2	3	4	5	
désapprouve	○	○	○	○	○	approuve

La traduction me permet de mieux suivre la matière présentée

	1	2	3	4	5	
désapprouve	○	○	○	○	○	approuve

Question sur le contenu

A combien de questions-clés doit-on trouver une réponse pour savoir comment une compagnie se différencie de la compétition? *

A.5.2 Feedback form for overall experience

Overall Feedback

* Required

A quelle date avez-vous suivi le webinaire? *

dd/mm/yyyy

Quelle est votre langue maternelle?

Quel niveau d'Anglais pensez vous avoir? *

	mauvais - Je me sens mal à l'aise	moyen - Je me débrouille	bon - Je me sens bien
En lecture	○	○	○
Ecoute passive	○	○	○
Expression orale	○	○	○
Expression écrite	○	○	○

Etes vous homme ou femme? *

- ○ homme
- ○ femme

Quel est votre âge? *

- ○ <18
- ○ 18-25

Depuis combien d'années faites-vous des études *

Evaluation Systeme

Impression générale

Le service est

	1	2	3	4	5	
affreux	○	○	○	○	○	super

L'expérience est

	1	2	3	4	5	
frustrante	○	○	○	○	○	excellente

Le système est

	1	2	3	4	5	
pas utile	○	○	○	○	○	utile

Utilité perçue

Un tel système me permet de mieux comprendre la matière présentée

	1	2	3	4	5	
désapprouve	○	○	○	○	○	approuve

Un tel système me permet de mieux suivre la matière présentée

	1	2	3	4	5	
désapprouve	○	○	○	○	○	approuve


Facilité d'usage perçue
J'ai aimé
1 2 3 4 5
désapprouve
approuve
Il y a tous les features dont on a besoin
1 2 3 4 5
désapprouve
approuve
Bon layout
1 2 3 4 5
désapprouve
approuve

Report on the 11th IWSLT Evaluation Campaign, IWSLT 2014

Mauro Cettolo[(1)] Jan Niehues[(2)] Sebastian Stüker[(2)] Luisa Bentivogli[(1)] Marcello Federico[(1)]

[(1)] FBK - Via Sommarive 18, 38123 Trento, Italy
[(2)] KIT - Adenauerring 2, 76131 Karlsruhe, Germany

Abstract

The paper overviews the 11th evaluation campaign organized by the IWSLT workshop. The 2014 evaluation offered multiple tracks on lecture transcription and translation based on the TED Talks corpus. In particular, this year IWSLT included three automatic speech recognition tracks, on English, German and Italian, five speech translation tracks, from English to French, English to German, German to English, English to Italian, and Italian to English, and five text translation track, also from English to French, English to German, German to English, English to Italian, and Italian to English. In addition to the official tracks, speech and text translation optional tracks were offered, globally involving 12 other languages: Arabic, Spanish, Portuguese (B), Hebrew, Chinese, Polish, Persian, Slovenian, Turkish, Dutch, Romanian, Russian. Overall, 21 teams participated in the evaluation, for a total of 76 primary runs submitted. Participants were also asked to submit runs on the 2013 test set (progress test set), in order to measure the progress of systems with respect to the previous year. All runs were evaluated with objective metrics, and submissions for two of the official text translation tracks were also evaluated with human post-editing.

1. Introduction

This paper overviews the results of the 2014 evaluation campaign organized by the International Workshop of Spoken Language Translation. The IWSLT evaluation has been running now for over a decade and has offered along these years a variety of speech translation tasks [1, 2, 3, 4, 5, 6, 7, 8, 9, 10]. The 2014 IWSLT evaluation continued along the line set in 2010, by focusing on the translation of TED Talks, a collection of public speeches covering many different topics. As in the previous two years, the evaluation included tracks for all the core technologies involved in the spoken language translation task, namely:

- Automatic speech recognition (ASR), i.e. the conversion of a speech signal into a transcript,
- Spoken language translation (SLT), that addressed the conversion and translation of a speech signal into a transcript in another language,
- Machine translation (MT), i.e. the translation of a polished transcript into another language.

However, with respect to previous rounds, new languages have been added to each track. The ASR track that previously included German and English, was extended by Italian. The SLT and MT track offered official English-French, English-German, German-English, English-Italian, and Italian-English translation directions. Besides the official evaluation tracks, many other optional translation directions were also offered. Optional SLT directions were English-Arabic and English-Chinese. Optional MT translation directions were: English from/to Arabic, Spanish, Portuguese (B), Hebrew, Chinese, Polish, Persian, Slovenian, Turkish, Dutch, Romanian, and Russian. For each official and optional translation direction, training and development data were supplied by the organizers through the workshop's website. Major parallel collections made available to the participants were the WIT^3 [11] corpus of TED talks, all data from the WMT 2014 workshop [12], the MULTIUN corpus, and the SETimes parallel corpus. A list of monolingual resources was provided too, that includes both freely available corpora and corpora available from LDC. Test data were released at the beginning of each test period, requiring participants to return one primary run and optional contrastive runs within one week. The schedule of the evaluation was organized as follows: June 2, release of training data; Sept 1–10, ASR test period; Sept 16–25, SLT test period (official directions); Sept 26–Oct 5, MT test period (official directions); Oct 6–17, MT and SLT test period of all optional directions.

All runs submitted by participants were evaluated with automatic metrics. In addition, manual evaluation was carried out for two MT tracks, namely the English-French and English-German tracks. Following the methodology introduced last year, systems were evaluated by calculating HTER values on post-edits created by professional translators. The rational behind this evaluation is to assess the utility of an MT output by measuring the post-editing effort needed by a professional translator to fix it.

This year, 21 sites participated (see Table 1) submitting a total of 76 primary runs: 15 to the ASR track, 16 to the SLT track, and 45 to the MT track (see Sections 3.3, 4.3, 5.3 for details).

In the rest of the paper we first outline the main goals of the IWSLT evaluation and then each single track in detail, in particular: its specifications, supplied language resources, evaluation methods, and results. The paper ends with some concluding remarks about the experiences gained in this eval-

Proceedings of the 11th International Workshop on Spoken Language Translation
Lake Tahoe, December 4th and 5th, 2014

uation exercise, followed by appendixes that complement the information given in the specific sections.

2. TED Talks

2.1. TED events

The translation of TED talks was introduced for the first time at IWSLT 2010. TED is a nonprofit organization that "invites the world's most fascinating thinkers and doers [...] to give the talk of their lives". Its website[1] makes the video recordings of the best TED talks available under the Creative Commons license. All talks have English captions, which have also been translated into many languages by volunteers worldwide. In addition to the official TED events held in North America, a series of independent TEDx events are regularly held around the world, which share the same format of the original TED talks but are hold in the language of the hosting country. Recently, an effort was made to set up a web repository [11] that distributes dumps of the available TED talks transcripts and translations under form of parallel texts, ready to use for training and evaluating MT systems.

Besides representing a popular benchmark for spoken language technology, the TED Talks task embeds interesting research challenges which are unique among the available speech recognition and machine translation benchmarks. TED Talks is a collection of rather short speeches (max 18 minutes each, roughly equivalent to 2,500 words) which cover a wide variety of topics. Each talk is delivered in a brilliant and original style by a very skilled speaker and, while addressing a wide audience, it pursues the goal of both entertaining and persuading the listeners on a specific idea. From the point of view of ASR, TED talks require copying with background noise – e.g. applauses and laughs by the public –, different accents including non native speakers, varying speaking rates, prosodic aspects, and, finally, narrow topics and personal language styles. From an application perspective, TED Talks transcription is the typical life captioning scenario, which requires producing polished subtitles in real-time.

From the point of view of machine translation, translating TED Talks implies dealing with spoken rather than written language, which is hence expected to be structurally less complex, formal and fluent. Moreover, as human translations of the talks are required to follow the structure and rythm of the English captions,[2] a lower amount of rephrasing and reordering is expected than in ordinary translation of written documents.

From an application perspective, TED Talks suggest translation tasks ranging from off-line translation of written captions, up to on-line speech translation, requiring a tight integration of MT with ASR possibly handling stream-based processing.

[1] http://www.ted.com

[2] See recommendations to translators in http://translations.ted.org/wiki.

3. ASR Track

3.1. Definition

The goal of the *Automatic Speech Recognition* (ASR) track for IWSLT 2014 was to transcribe English TED talks, as well as German and Italian TEDx talks. The speech in TED lectures is in general planned, well articulated, and recorded in high quality. The main challenges for ASR in these talks are to cope with a large variability of topics, the presence of non-native speakers, and the rather informal speaking style. For the TEDx talks the recording conditions are a little bit more difficult than for the English TED talks. While the TEDx talks aim to mimic the TED talks, they are not as well prepared and well rehearsed as the TED lectures, and recording is often done by amateurs resulting in often poorer recording quality than for the TED lectures.

The result of the recognition of the talks is used for two purposes. It is used to measure the performance of ASR systems on the talks and it is used as input for the spoken language translation evaluation (SLT), see Section 4.

3.2. Evaluation

Participants had to submit the results of the recognition of the tst2014 set in CTM format. The word error rate was measured case-insensitive. After the end of the evaluation a preliminary scoring was performed with the first set of references. This was followed by an adjudication phase in which participants could point out errors in the reference transcripts. The adjudication results were collected and combined into the final set of references with which the official scores were calculated.

In order to measure the progress of the systems over the years on English and German, participants also had to provide results on the test set from 2013, i.e. *tst2013*.

3.3. Submissions

For this year's evaluation we received primary submissions from eight sites as well as one combined submission by the EU-BRIDGE project. Seven sites participated in the English evaluation, three sites in the German evaluation and four sites in the Italian one. For English we further received a total of seven contrastive submissions from five sites. For German we received three contrastive submissions from one participant. For Italian we receieved five contrastive submissions from three sites. Also, for English we received a joint submission by the project EU-BRIDGE which was a ROVER combination of the partners' outputs and for which no separate system description was submitted.

3.4. Results

The detailed results of the primary submissions of the evaluation in terms of word error rate (WER) can be found in Appendix A.1. The word error rate of the submitted systems in in the range of 8.4%–19.7% for English, 24.0%–38.8% for

Table 1: List of Participants

EU-BRIDGE	RWTH& UEDIN& KIT& FBK[13]
FBK	Fondazione Bruno Kessler, Italy [14, 15]
HKUST	Hong Kong University of Science and Technology, Hong Kong [16]
IOIT	Inst. of Inform. and Techn., Vietn. Acad. of Science and Techn. & Thai Nguyen University, Vietnam[17]
KIT	Karlsruhe Institute of Technology, Germany [18, 19]
KLE	Pohang University of Science and Technology, Republic of Korea
LIA	Laboratoire Informatique d'Avignon (LIA) University of Avignon, France [20]
LIMSI	LIMSI - LIMSI, France [21]
LIUM	LIUM, University of Le Mans, France [22]
MIRACL	MIRACL Laboratory Pôle Technologique, Tunisia & LORIA Nancy, France [23]
MITLL-AFRL	Mass. Institute of Technology/Air Force Research Lab., USA
NICT	National Institute of Communications Technology, Japan [24, 25]
NTT-NAIST	NTT Communication Science Labs, Japan & NAIST[26]
PJIIT	Polish-Japanese Institute of Information Technology, Poland [27]
RWTH	Rheinisch-Westfälische Technische Hochschule Aachen, Germany [28]
SFAX	Sfax University, Tunisia
UEDIN	University of Edinburgh, United Kingdom [29, 30]
UMONTREAL	Université de Montréal, Canada
USFD	University of Sheffield, United Kingdom [31]
USTC	National Engineering Laboratory of Speech and Lang. Inform. Proc., Univ. of Science and Techn. of China [32]
VECSYS-LIUM	Vecsys Technologies, France & University of Le Mans, France [22]

German, and 21.9%–25.4% for Italian.

In German, the fact that TEDx have sometimes worse recording conditions than TED talks was reflected by the fact that two talks in the German tst2014 had WERs above 40%. WERs for all other talks were in the range from 9% to 32%.

For English, it can be seen that all participants from IWSLT 2013 made progress, many significant progress, e.g., bringing down the WER from 13.5% to 10.6% on *tst2013*, a relative reduction of 21% over the course of one year. For German, the best performing system only made minor progress, while one of the runner-ups made significant progress and one participant essentially stood the same.

4. SLT Track

4.1. Definition

The SLT track required participants to translate the English, German and Italian talks of *tst2014* from the audio signal (see Section 3). The challenge of this translation task over the MT track is the necessity to deal with automatic, and in general error prone, transcriptions of the audio signal, instead of correct human transcriptions.

For German and Italian, participants had to translate into English. For English as source language, participants had to translate into French. In addition, participants could also optionally translate from English into one of the following languages: German, Italian, Arabic and Mandarin Chinese.

4.2. Evaluation

For the evaluation, participants could choose to either use their own ASR technology, or to use ASR output provided by the conference organizers. In order to facilitate scoring, participants had to segment the audio according to the manual reference segmentation provided by the organizers of the evaluation.

For English, the ASR output provided by the organizers was a ROVER combination of the output from five submissions to the ASR track. The result of the ROVER had a WER of 8.2%. For German and Italian we used the two single best scored submissions, as ROVER combination with other systems did not give any performance gains.

The results of the translation had to be submitted in the same format as for the machine translation track (see Section 5).

4.3. Submissions

We received 16 primary and 31 contrastive submissions from nine participants, English to French receiving the most submissions.

4.4. Results

The detailed results of the automatic evaluation in terms of BLEU and TER can be found in Appendix A.1.

Table 2: Monolingual resources for official language pairs

data set	lang	sent	token	voc
train	De	183k	3.36M	124.7k
	En	188k	3.81M	63.4k
	Fr	186k	4.00M	77.0k
	It	185k	3.49M	90.2k

5. MT Track

5.1. Definition

The MT TED track basically corresponds to a subtitling translation task. The natural translation unit considered by the human translators volunteering for TED is indeed the single caption — as defined by the original transcript — which in general does not correspond to a sentence, but to fragments of it that fit the caption space. While translators can look at the context of the single captions, arranging the MT task in this way would make it particularly difficult, especially when word re-ordering across consecutive captions occurs. For this reason, we preprocessed all the parallel texts to re-build the original sentences, thus simplifying the MT task.

For each official and optional translation direction, in-domain training and development data were supplied through the website of WIT[3] [11], while out-of-domain training data through the workshop's website. As usual, some of the talks added to the TED repository during the last year have been used to define the new evaluation sets (*tst2014*), while the remaining new talks have been included in the training sets. For reliably assessing progress of MT systems over the years, the evaluation sets *tst2013* of edition 2013 were distributed together with *tst2014* as progressive test sets, when available. Development sets (*dev2010, tst2010, tst2011* and *tst2012*) are either the same of past editions or, in case of new language pairs, have been built upon the same talks.

Evaluation sets *tst2014* of $DeEn$ and $ItEn$ MT tasks derive from those prepared for ASR/SLT tracks, which consist of TEDx talks delivered in German and Italian language, respectively; therefore, no overlap exists with any other TED talk involved in other tasks. Since the $DeEn$ TEDx based MT task was proposed in 2013 as well, the *tst2013* has been released as progressive test set; on the contrary, it is the first time that Italian is involved in ASR/SLT tracks, therefore no evaluation set is available for assessing progress. A single TEDx based development set was released for each pair, together with standard TED based development sets *dev2010, tst2010, tst2011* and *tst2012* sets.

Tables 2 and 3 provides statistics on in-domain texts supplied for training, development and evaluation purposes for the official directions.

MT baselines were trained from TED data only, i.e. no additional out-of-domain resources were used. The standard tokenization via the tokenizer script released with the Europarl corpus [33] was applied to all languages, with the exception of Chinese and Arabic languages, which were preprocessed by, respectively: the Stanford Chinese Segmenter [34] and the QCRI-normalizer.[3]

Table 3: Bilingual resources for official language pairs.

MT task	set	sent	tokens		talks
En→Fr			En	Fr	
	train	179k	3.63M	3.88M	1415
	TED.dev2010	887	20,1k	20,2k	8
	TED.tst2010	1,664	32,0k	33,9k	11
	TED.tst2011	818	14,5k	15,6k	8
	TED.tst2012	1,124	21,5k	23,5k	11
	TED.tst2013	1,026	21,7k	23,3k	16
	TED.tst2014	1,305	24,8k	27,5k	15
En↔De			En	De	
	train	172k	3.46M	3.24M	1361
	TED.dev2010	887	20,1k	19,1k	8
	TED.tst2010	1,565	32,0k	30,3k	11
	TED.tst2011	1,433	26,9k	26,3k	16
	TED.tst2012	1,700	30,7k	29,2k	15
→	TED.tst2013	993	20,9k	19,7k	16
	TED.tst2014	1,305	24,8k	23,8k	15
←	TEDx.dev2012	1,165	21,6k	20,8k	7
	TEDx.tst2013	1,363	23,3k	22,4k	9
	TEDx.tst2014	1,414	28,1k	27,6k	10
En↔It			En	It	
	train	182k	3.68M	3.44M	1434
	TED.dev2010	887	20,1k	17,9k	8
	TED.tst2010	1,529	31,0k	28,7k	10
	TED.tst2011	1,433	26,9k	24,5k	16
	TED.tst2012	1,704	30,7k	28,2k	15
→	TED.tst2013	1,402	30,1k	28,7k	21
	TED.tst2014	1,183	22,6k	21,2k	14
←	TEDx.dev2014	1,056	28,9k	28,6k	13
	TEDx.tst2014	883	25,9k	26,5k	13

The baselines were developed with the Moses toolkit. Translation and lexicalized reordering models were trained on the parallel training data; 5-gram LMs with improved Kneser-Ney smoothing were estimated on the target side of the training parallel data with the IRSTLM toolkit. The weights of the log-linear interpolation model were optimized with the MERT procedure provided with Moses, mostly on the development sets *tst2010*; the exceptions are: TEDx tasks, where the TEDx based development sets were used; the two pairs involving Slovenian, where *dev2012* were employed.

5.2. Evaluation

The participants to the MT track had to provide the results of the translation of the test sets in NIST XML format. The output had to be case-sensitive and had to contain punctuation

[3] QCRI-normalizer was specifically developed for IWSLT Evaluation Campaigns by P. Nakov and F. Al-Obaidli at Qatar Computing Research Institute.

(case+punc).

The quality of the translations was measured automatically against the human translations created by the TED open translation project, and by human subjective evaluation (Section 5.5). Tokenization scripts were applied automatically to all run submissions prior to evaluation.

Evaluation scores were calculated for the two automatic standard metrics BLEU and TER, as implemented in mteval-v13a.pl[4] and tercom-0.7.25[5], respectively.

5.3. Submissions

We received submissions from 14 different sites. On official pairs, the total number of primary runs is 39: 20 on *tst2014* and 19 on *tst2013*; 15 primary runs regard the $EnFr$ pair, 10 the $EnDe$ and 14 the $DeEn$; in addition, we were asked to evaluate also 64 contrastive runs.

Concerning the optional pairs, we received 48 primary runs (25 on *tst2014* and 23 on *tst2013*) and 20 contrastive submissions. The tasks that attracted the most interest are those involving Chinese: 8 primary runs were submitted for $EnZh$, 8 for $ZhEn$. The other submissions involve Arabic, Polish, Farsi, Hebrew, Turkish and Slovenian.

5.4. Results

Table 4: BLEU and TER scores of baseline SMT systems on all *tst2014* sets. (†) TEDx test set. (*) Char-level scores.

pair		direction → BLEU	direction → TER	direction ← BLEU	direction ← TER
En	Fr	32.07	48.62	–	–
	De	18.33	62.11	†17.89	†64.91
	It	27.15	53.19	†26.12	†55.30
	Ar	11.13	73.01	20.59	62.62
	Es	31.31	48.29	33.88	45.96
	Fa	11.31	71.20	16.74	72.02
	He	15.91	65.62	24.41	58.38
	Nl	22.77	58.38	27.82	52.98
	Pl	9.63	82.81	14.28	68.96
	Pt	31.25	47.25	36.44	42.80
	Ro	18.05	65.25	25.06	54.62
	Ru	11.74	71.99	15.91	69.73
	Sl	8.46	73.94	14.27	71.03
	Tr	7.75	78.69	12.88	77.15
	Zh	*16.49	*79.50	11.74	72.31

First of all, for reference purposes Table 4 shows BLEU and TER scores on the *tst2014* evaluation sets of the baseline systems we developed as described in Section 5.1.

The results on the official test set for each participant are shown in Appendix A.1. For most languages, we show the case-sensitive and case-insensitive BLEU and TER scores. In contrast to the other language pairs, for English to Chinese character-level scores are reported.

These results also show again the scores of the baseline system. Thereby, it is possible to see the improvements of the submitted systems on the different languages over the baseline system.

In Appendix A.2 the results on the progress test sets *test2013* are shown. When comparing the results to the submissions from last year, the performance could be improved in nearly all tasks.

5.5. Human Evaluation

Human evaluation was carried out on primary runs submitted by participants to two of the official MT TED tracks, namely the MT English-German ($EnDe$) track and MT English-French ($EnFr$) track. Following the methodology introduced last year, human evaluation was based on *Post-Editing*, and HTER (Human-mediated Translation Edit Rate) was adopted as the official evaluation metric to rank the systems.

Post-Editing, i.e. the manual correction of machine translation output, has long been investigated by the translation industry as a form of machine assistance to reduce the costs of human translation. Nowadays, Computer-aided translation (CAT) tools incorporate post-editing functionalities, and a number of studies [35, 36] demonstrate the usefulness of MT to increase professional translators' productivity. The MT TED task offered in IWSLT can be seen as an interesting application scenario to test the utility of MT systems in a real subtitling task.

From the point of view of the evaluation campaign, our goal was to adopt a human evaluation framework able to maximize the benefit to the research community, both in terms of information about MT systems and data and resources to be reused. With respect to other types of human assessment, such as judgments of translation quality (i.e. adequacy/fluency and ranking tasks), the post-editing task has the double advantage of producing *(i)* a set of edits pointing to specific translation errors, and *(ii)* a set of additional reference translations. Both these byproducts are very useful for MT system development and evaluation. Furthermore, HTER[37] - which consists of measuring the minimum edit distance between the machine translation and its manually post-edited version - has been shown to correlate quite well with human judgments of MT quality.

The human evaluation setup and the collection of post-editing data are presented in Section 5.5.1, whereas the results of the evaluation are presented in Section 5.5.2.

5.5.1. Evaluation Setup and Data Collection

The human evaluation (HE) dataset created for each MT track was a subset of the corresponding 2013 progress test

[4] http://www.itl.nist.gov/iad/mig/tests/mt/2009/

[5] http://www.cs.umd.edu/ snover/tercom/

set *(tst2013)*.[6] Both the $EnDe$ and $EnFr$ *tst2013* datasets are composed of 16 TED Talks, and we selected around the initial 60% of each talk. This choice of selecting a consecutive block of sentences for each talk was determined by the need of realistically simulating a caption post-editing task on several TED talks. The resulting HE sets are composed of 628 segments for $EnDe$ and 622 segments for $EnFr$, both corresponding to around 11,000 words.

In order to evaluate the MT systems, the *bilingual* post-editing task was chosen, where professional translators are required to post-edit the MT output directly according to the source sentence. Bilingual post-editing is expected to give more accurate results than monolingual post-editing as post-editors do not depend on an given - and possibly imprecise - translation. Then, HTER scores were calculated on the created post-edits. HTER [37] is a semi-automatic metric derived from TER (Translation Edit Rate). TER measures the amount of editing that a human would have to perform to change a machine translation so that it exactly matches a given reference translation. HTER is a variant of TER where a new reference translation is generated by applying the minimum number of post-edits to the given MT output. This new *targeted* reference is then used as the only reference translation to calculate the TER of the MT output.

An interesting outcome of last year's manual evaluation [10] was that the most informative and reliable HTER was not obtained by using only the targeted reference but by exploiting all the post-edits of the evaluated MT outputs. According to these results, also this year systems were officially ranked according to HTER calculated on multiple references.

As for the systems to be evaluated, this year we received five primary runs for the $EnDe$ track and seven for the $EnFr$ track. All the five $EnDe$ MT outputs were post-edited, whereas for the $EnFr$ track we decided to post-edit only five MT outputs out of the seven received. This reduction is not supposed to affect the official evaluation results - since all the participating systems are evaluated with HTER based on multiple post-edits - and it allowed us to respect the budget limitations while offering the community five additional reference translations for a high number of segments (around 60% of the test sets) and for two different language pairs. The five MT outputs selected for post-editing in the $EnFr$ task are the top-5 ranked systems according to automatic evaluation (see Appendix A).

In the preparation of the post-editing data to be collected, some constraints were identified to ensure the soundness of the evaluation: *(i)* each translator must post-edit all segments of the HE set, *(ii)* each translator must post-edit the segments of the HE set only once, and *(iii)* each MT system must be equally post-edited by all translators. Furthermore, in order to cope with the variability of post-editors (i.e. some translators could systematically post-edit more than others) we devised a scheme that dispatches MT outputs to translators both randomly and satisfying the uniform assignment constraints. For each task, five documents were hence prepared including all source segments of the HE set and, for each source segment, one MT output selected from one of the five systems.

Documents were delivered to a language service provider together with instructions to be passed on to the translators, and the post-editing tasks were run using an enterprise-level CAT tool developed under the MateCat project[7]. Both the post-editing interface and the guidelines given to translators are presented in Appendix B.

For each task, the resulting collected data consist of five new reference translations for each of the sentences of the HE set. Each one of these five references represents the targeted translation of the system output from which it was derived. From the point of view of the system output, one targeted translation and other four translations are available.

The main characteristics of the work carried out by post-editors are presented in Table 5 for the $EnDe$ task and in Table 6 for the $EnFr$ task, and largely confirm last year's findings. In the tables, the post-editing effort for each translator is given. Post-editing effort is to be interpreted as the number of actual edit operations performed to produce the post-edited version and - consequently - it is calculated as the HTER of all the system sentences post-edited by each single translator. It is interesting to see that the PE effort is similar for both language pairs, and also highly variable among post-editors, ranging from 19.69% to 40.91% for the $EnDe$ task, and from 17.47% to 39.65% for the $EnFr$ task. Data about weighted standard deviation confirm post-editor variability, showing that the five translators produced quite different post-editing effort distributions.

Table 5: En-De task: Post-editing information for each Post-editor

PEditor	PE Effort	*std-dev*	Sys TER	*std-dev*
PE 1	32.17	18.80	56.05	20.23
PE 2	19.69	13.56	56.32	20.34
PE 3	40.91	17.23	56.18	19.58
PE 4	27.56	14.71	55.93	20.02
PE 5	24.99	15.62	55.63	19.88

Table 6: En-Fr task: Post-editing information for each Post-editor

PEditor	PE Effort	*std-dev*	Sys TER	*std-dev*
PE 1	34.96	20.21	42.60	17.61
PE 2	17.47	14.76	42.81	17.98
PE 3	23.68	14.17	43.02	17.74
PE 4	39.65	20.47	42.27	17.78
PE 5	19.73	14.07	42.86	17.72

[6] Since all the data produced for human evaluation will be made publicly available thorough the WIT^3 repository, we used the 2013 test set in order to keep the 2014 test set blind to be used as a progress test for next year's evaluation.

[7] www.matecat.com

To further study post-editor variability, we exploited the official reference translations available for the two TED tracks and we calculated the TER of the MT outputs assigned to each translator for post-editing ("Sys TER" Column in Tables 5 and 6), as well as the related standard deviation.

As we can see from the tables, the documents presented to translators (composed of segments produced by different systems) are very homogeneous, as they show very similar TER scores and standard deviation figures. This also confirms that the procedure followed in data preparation was effective.

The variability observed in post-editing effort - despite the similarity of the input documents - is most probably due to translators' subjectivity in carrying out the post-editing task. Thus, post-editor variability is an issue to be addressed to ensure a sound evaluation of the systems.

5.5.2. *Evaluation Results*

As anticipated above, last year's human evaluation results demonstrated that HTER computed against all the references produced by all post-editors allowed a more reliable and consistent evaluation of MT systems with respect to HTER calculated against the targeted reference only. Indeed, the HTER reduction obtained using all post-edits clearly showed that exploiting all the available reference translations is a viable way to control and overcome post-editors' variability. For this reason, also this year systems were officially ranked according to HTER calculated on multiple references.

For the $EnDe$ task, HTER was calculated using all the five post-edits available, i.e. for each system the targeted translation and the additional four references were used. For the $EnFr$ task, since the post-edits for two MT outputs had not been created, in order to avoid biases only four post-edits out of five were used to calculate HTER, namely excluding from each system's evaluation its targeted translation.

The official results of human evaluation are given in Tables 7 and 8, which also present a comparison of HTER scores and rankings with TER results - on the HE set and on the full test set - calculated against the official reference translation used for automatic evaluation (see Section 5.2).[8] For the $EnFr$ task, the official HTER results presented in Table 8 for FBK and MIRACL (which do not have a corresponding post-edit) are those obtained on the combination of the four post-edits which gave the best results.

In general, the very low HTER results obtained in both tasks demonstrate that the overall quality of the systems is very high. Moreover, all systems are very close to each other. To establish the reliability of system ranking, for all pairs of systems we calculated the statistical significance of the observed differences in performance. Statistical significance was assessed with the *approximate randomization* method [38], a statistical test well-established in the NLP community [39] and that, especially for the purpose of MT evaluation, has been shown [40] to be less prone to type-I errors than the bootstrap method [41]. The approximate randomization test was based on 10,000 iterations, and differences were considered statistically significant at $p < 0.01$. According to this test, for both tasks a winning system cannot be indicated, as there is no system that is significantly better than all other systems. In particular, for the $EnDe$ task only the bottom-ranking system (KLE) is significantly worse than all the other systems. For the $EnFr$ task, in Table 8 we report - next to the HTER score of each system - the name of the first system in the ranking with respect to which differences are statistically significant. We can see that only the two top-ranking systems are significantly better than the four bottom-ranking systems (from UEDIN to MIRACL), whereas all the other systems significantly differ only with respect to MIRACL.

Furthermore, for comparison purposes, Table 8 presents additional HTER results calculated on all the five post-edits available for the $EnFr$ task. First, it is interesting to note the further HTER reduction achieved, especially for the five top-scoring systems since their corresponding targeted reference was added. Also, comparing the two language pairs, we see that the HTER scores obtained for $EnFr$ with five reference translations are overall lower than those obtained for $EnDe$, indicating that systems translating into French perform better than systems translating into German.

A number of additional observations can be drawn by comparing the official HTER results with TER results. In general, for both tasks we can see that HTER reduces the edit rate of more than 50% with respect to TER. Moreover,

Table 7: En-De Task: Official human evaluation results

System Ranking	**HTER** ***HE Set 5 PErefs***	**TER** HE Set ref	**TER** Test Set ref
EU-BRIDGE	**19.22**	54.55	53.62
UEDIN	**19.93**	56.32	55.12
KIT	**20.88**	54.88	53.83
NTT-NAIST	**21.32**	54.68	53.86
KLE	**28.75**	59.67	58.27
Rank Corr.		0.60	0.70

Table 8: En-Fr Task: Official human evaluation results

System Ranking	**HTER** ***HE Set 4 PErefs***	**HTER** HE Set 5 PErefs	**TER** HE Set ref	**TER** Test Set ref
EU-BRIDGE	**19.21**$^{\text{UEDIN}}$	16.48	42.64	43.27
RWTH	**19.27**$^{\text{UEDIN}}$	16.55	41.82	42.58
KIT	**20.89**$^{\text{MIRACL}}$	17.64	42.33	43.09
UEDIN	**21.52**$^{\text{MIRACL}}$	17.23	43.28	43.80
MITLL-AFRL	**22.64**$^{\text{MIRACL}}$	18.69	43.48	44.05
FBK	**22.90**$^{\text{MIRACL}}$	22.29	44.28	44.83
MIRACL	**33.61**	32.90	52.19	51.96
Rank Corr.		0.96	0.90	0.90

[8] Note that since HTER and TER are edit-distance measures, lower numbers indicate better performances

the correlation between evaluation metrics is measured using *Spearman's rank correlation coefficient* $\rho \in$ [-1.0, 1.0], with ρ = 1.0 if all systems are ranked in same order, ρ = -1.0 if all systems ranked in reverse order and ρ = 0.0 if no correlation exists. We can see from the tables that TER rankings correlate well with the official HTER.

To conclude, the post-editing task introduced this year for manual evaluation brought benefit to the IWSLT community, and in general to the MT field. In fact, producing post-edited versions of the participating systems' outputs allowed us to carry out a quite informative evaluation by minimizing the variability of post-editors, who naturally tend to diverge from the post-editing guidelines and personalize their translations. Moreover, a number of additional reference translations will be available for further development and evaluation of MT systems.

6. Conclusions

We have reported on the evaluation campaign organized for the eleventh edition of the IWSLT workshop. The evaluation has addressed three tracks: automatic speech recognition of talks (in English, German, and Italian), speech-to-text translation, and text-to-text translation, both from German to English, English to German, and English to French. Besides the official translation directions, many optional translation tasks were available, too, including 12 additional languages. For each task, systems had to submit runs on three different test sets: a newly created official test set, and a progress test set created and used for the 2013 evaluation. This year, 21 participants took part in the evaluation, submitting a total of 76 primary runs, which were all scored with automatic metrics. We also manually evaluated runs of the English-German and English-French text translation tracks. In particular, we asked professional translators to post-edit system outputs on a subset of the 2013 progress test set, in order to produce *close references* for them. While we have observed a significant variability among translators, in terms of post-edit effort, we could obtain more reliable scores by using all the produced post-edits as reference translations. By using the HTER metric, for both tracks the post-edit effort of the best performing system results remarkably low, namely around 19%. Considering that this is still an upper bound of the ideal HTER score, this percentage of post-editing seems to be another strong argument supporting the utility of machine translation for human translators.

7. Acknowledgements

Research Group 3-01' received financial support by the '*Concept for the Future*' of Karlsruhe Institute of Technology within the framework of the German Excellence Initiative. The work leading to these results has received funding from the European Union under grant agreement no 287658 — Bridges Across the Language Divide (EU-BRIDGE).

8. References

[1] Y. Akiba, M. Federico, N. Kando, H. Nakaiwa, M. Paul, and J. Tsujii, "Overview of the IWSLT04 Evaluation Campaign," in *Proceedings of the International Workshop on Spoken Language Translation*, Kyoto, Japan, 2004, pp. 1–12.

[2] M. Eck and C. Hori, "Overview of the IWSLT 2005 evaluation campaign," in *Proceedings of the International Workshop on Spoken Language Translation*, Pittsburgh, PA, 2005, pp. 1–22.

[3] P. Michael, "Overview of the IWSLT 2006 Evaluation Campaign," in *Proceedings of the International Workshop on Spoken Language Translation*, Kyoto, Japan, 2006, pp. 1–15.

[4] C. S. Fordyce, "Overview of the IWSLT 2007 evaluation campaign," in *Proceedings of the International Workshop on Spoken Language Translation*, Trento, Italy, 2007, pp. 1–12.

[5] M. Paul, "Overview of the IWSLT 2008 Evaluation Campaign," in *Proceedings of the International Workshop on Spoken Language Translation*, Waikiki, Hawaii, 2008, pp. 1–17.

[6] ——, "Overview of the IWSLT 2009 Evaluation Campaign," in *Proceedings of the sixth International Workshop on Spoken Language Translation*, Tokyo, Japan, 2009, pp. 1–18.

[7] M. Paul, M. Federico, and S. Stüker, "Overview of the IWSLT 2010 Evaluation Campaign," in *Proceedings of the International Workshop on Spoken Language Translation*, Paris, France, 2010, pp. 3–27.

[8] M. Federico, L. Bentivogli, M. Paul, and S. Stüker, "Overview of the IWSLT 2011 Evaluation Campaign," in *Proceedings of the International Workshop on Spoken Language Translation*, San Francisco, USA, 2011, pp. 11–27.

[9] M. Federico, M. Cettolo, L. Bentivogli, M. Paul, and S. Stüker, "Overview of the IWSLT 2012 Evaluation Campaign," in *Proceedings of the International Workshop on Spoken Language Translation*, Hong Kong, HK, 2012, pp. 11–27.

[10] M. Cettolo, J. Niehues, S. Stüker, L. Bentivogli, and M. Federico, "Report on the 10th IWSLT Evaluation Campaign," in *Proceedings of the Tenth International Workshop on Spoken Language Translation (IWSLT 2013)*, Heidelberg, Germany, 2013.

[11] M. Cettolo, C. Girardi, and M. Federico, "WIT^3: Web Inventory of Transcribed and Translated Talks," in *Proceedings of the Annual Conference of the European Association for Machine Translation*

(EAMT), Trento, Italy, May 2012. [Online]. Available: http://hltshare.fbk.eu/EAMT2012/html/Papers/59.pdf

[12] O. Bojar, C. Buck, C. Federmann, B. Haddow, P. Koehn, J. Leveling, C. Monz, P. Pecina, M. Post, H. Saint-Amand, R. Soricut, L. Specia, and A. Tamchyna, "Findings of the 2014 Workshop on Statistical Machine Translation," in *Proceedings of the Ninth Workshop on Statistical Machine Translation*, Baltimore, MD, USA, 2014.

[13] M. Freitag, J. Wuebker, S. Peitz, H. Ney, M. Huck, A. Birch, N. Durrani, P. Koehn, M. Mediani, I. Slawik, J. Niehues, E. Cho, A. Waibel, N. Bertoldi, M. Cettolo, and M. Federico, "Combined Spoken Language Translation," in *Proceedings of the 11th International Workshop on Spoken Language Translation (IWSLT)*, Lake Tahoe, CA, 2014.

[14] B. Babaali, R. Serizel, S. Jalalvand, D. Falavigna, R. Gretter, and D. Giuliani, "FBK @ IWSLT 2014 - ASR track," in *Proceedings of the 11th International Workshop on Spoken Language Translation (IWSLT)*, Lake Tahoe, CA, 2014.

[15] N. Bertoldi, P. Mathur, N. Ruiz, and M. Federico, "FBK's Machine Translation and Speech Translation Systems for the IWSLT 2014 Evaluation Campaign," in *Proceedings of the 11th International Workshop on Spoken Language Translation (IWSLT)*, Lake Tahoe, CA, 2014.

[16] M. Beloucif, C.-K. Lo, and D. Wu, "Improving tuning against MEANT," in *Proceedings of the 11th International Workshop on Spoken Language Translation (IWSLT)*, Lake Tahoe, CA, 2014.

[17] Q. B. Nguyen, T. T. Vu, and C. M. Luong, "The Speech Recognition Systems of IOIT for IWSLT 2014," in *Proceedings of the 11th International Workshop on Spoken Language Translation (IWSLT)*, Lake Tahoe, CA, 2014.

[18] K. Kilgour, M. Heck, M. Müller, M. Sperber, S. Stüker, and A. Waibel, "The 2014 KIT IWSLT Speech-to-Text Systems for English, German and Italian," in *Proceedings of the 11th International Workshop on Spoken Language Translation (IWSLT)*, Lake Tahoe, CA, 2014.

[19] I. Slawik, M. Mediani, J. Niehues, Y. Zhang, E. Cho, T. Herrmann, T.-L. Ha, and A. Waibel, "The KIT Translation Systems for IWSLT 2014," in *Proceedings of the 11th International Workshop on Spoken Language Translation (IWSLT)*, Lake Tahoe, CA, 2014.

[20] M. Morchid, S. Huet, and R. Dufour, "A Topic-based Approach for Post-processing Correction of Automatic Translations," in *Proceedings of the 11th International Workshop on Spoken Language Translation (IWSLT)*, Lake Tahoe, CA, 2014.

[21] N. Segal, H. Bonneau-Maynard, Q. K. Do, A. Allauzen, J.-L. Gauvain, L. Lamel, and F. Yvon, "LIMSI English-French Speech Translation System," in *Proceedings of the 11th International Workshop on Spoken Language Translation (IWSLT)*, Lake Tahoe, CA, 2014.

[22] A. Rousseau, L. Barrault, P. Deléglise, Y. Estève, H. Schwenk, S. Bennacef, A. Muscariello, and S. Vanni, "The LIUM English-to-French Spoken Language Translation System and the Vecsys/LIUM Automatic Speech Recognition System for Italian Language for IWSLT 2014," in *Proceedings of the 11th International Workshop on Spoken Language Translation (IWSLT)*, Lake Tahoe, CA, 2014.

[23] A. B. Romdhane, S. Jamoussi, A. B. Hamadou, and K. Smaili, "Phrase-based Language Modelling for Statistical Machine Translation," in *Proceedings of the 11th International Workshop on Spoken Language Translation (IWSLT)*, Lake Tahoe, CA, 2014.

[24] P. Shen, X. Lu, X. Hu, N. Kanda, M. Saiko, and C. Hori, "The NICT ASR System for IWSLT 2014," in *Proceedings of the 11th International Workshop on Spoken Language Translation (IWSLT)*, Lake Tahoe, CA, 2014.

[25] X. Wang, A. Finch, M. Utiyama, T. Watanabe, and E. Sumita, "The NICT Translation System for IWSLT 2014," in *Proceedings of the 11th International Workshop on Spoken Language Translation (IWSLT)*, Lake Tahoe, CA, 2014.

[26] K. Sudoh, G. Neubig, K. Duh, and K. Hayashi, "NTT-NAIST Syntax-based SMT Systems for IWSLT 2014," in *Proceedings of the 11th International Workshop on Spoken Language Translation (IWSLT)*, Lake Tahoe, CA, 2014.

[27] K. Wolk and K. Marasek, "Polish - English Speech Statistical Machine Translation Systems for the IWSLT 2014," in *Proceedings of the 11th International Workshop on Spoken Language Translation (IWSLT)*, Lake Tahoe, CA, 2014.

[28] J. Wuebker, S. Peitz, A. Guta, and H. Ney, "The RWTH Aachen Machine Translation Systems for IWSLT 2014," in *Proceedings of the 11th International Workshop on Spoken Language Translation (IWSLT)*, Lake Tahoe, CA, 2014.

[29] P. Bell, P. Swietojanski, J. Driesen, M. Sinclair, F. McInnes, and S. Renals, "The UEDIN ASR Systems for the IWSLT 2014 Evaluation," in *Proceedings of the 11th International Workshop on Spoken Language Translation (IWSLT)*, Lake Tahoe, CA, 2014.

[30] A. Birch, M. Huck, N. Durrani, N. Bogoychev, and P. Koehn, "Edinburgh SLT and MT System Description for the IWSLT 2014 Evaluation," in *Proceedings of*

the 11th International Workshop on Spoken Language Translation (IWSLT), Lake Tahoe, CA, 2014.

[31] R. W. M. Ng, M. Doulaty, R. Doddipatla, W. Aziz, K. Shah, O. Saz, M. Hasan, G. Alharbi, L. Specia, and T. Hain, "The USFD SLT system for IWSLT 2014," in *Proceedings of the 11th International Workshop on Spoken Language Translation (IWSLT)*, Lake Tahoe, CA, 2014.

[32] S. Wang, Y. Wang, J. Li, Y. Cui, and L. Dai, "The USTC Machine Translation System for IWSLT 2014," in *Proceedings of the 11th International Workshop on Spoken Language Translation (IWSLT)*, Lake Tahoe, CA, 2014.

[33] P. Koehn, "Europarl: A parallel corpus for statistical machine translation," in *Proceedings of the Tenth Machine Translation Summit (MT Summit X)*, Phuket, Thailand, September 2005, pp. 79–86.

[34] H. Tseng, P. Chang, G. Andrew, D. Jurafsky, and C. Manning, "A conditional random field word segmenter," in *Fourth SIGHAN Workshop on Chinese Language Processing*, 2005.

[35] M. Federico, A. Cattelan, and M. Trombetti, "Measuring user productivity in machine translation enhanced computer assisted translation," in *Proceedings of the Tenth Conference of the Association for Machine Translation in the Americas (AMTA)*, 2012. [Online]. Available: http://www.mt-archive.info/AMTA-2012-Federico.pdf

[36] S. Green, J. Heer, and C. D. Manning, "The efficacy of human post-editing for language translation," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 2013, pp. 439–448.

[37] M. Snover, B. Dorr, R. Schwartz, L. Micciulla, and J. Makhoul, "A Study of Translation Edit Rate with Targeted Human Annotation," in *Proceedings of the The Seventh Conference of the Association for Machine Translation in the Americas (AMTA)*, Cambridge, USA, 2006, pp. 223–231.

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

[39] 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.

[40] S. Riezler and J. T. Maxwell, "On some pitfalls in automatic evaluation and significance testing for MT," in *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*. Ann Arbor, Michigan: Association for Computational Linguistics, June 2005, pp. 57–64. [Online]. Available: http://www.aclweb.org/anthology/W/W05/W05-0908

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

Appendix A. Automatic Evaluation

"case+punc" evaluation : case-sensitive, with punctuations tokenized
"no_case+no_punc" evaluation : case-insensitive, with punctuations removed

A.1. Official Testset (*tst2014*)

· All the sentence IDs in the IWSLT 2014 test set were used to calculate the automatic scores for each run submission.
· ASR and MT systems are ordered according to the *WER* and *BLEU* metrics, respectively.
· All automatic evaluation metric scores are given as percent figures (%).

TED : ASR English (ASR$_{EN}$)

System	WER	(# Errors)
NICT	**8.4**	**(1,831)**
EU-BRIDGE	9.8	(2,138)
MITLL-AFRL	9.9	(2,153)
KIT	11.4	(2,475)
FBK	11.4	(2,492)
LIUM	12.3	(2,689)
UEDIN	12.7	(2,763)
IOIT	19.7	(4,283)

TED : ASR German (ASR$_{DE}$)

System	WER	(# Errors)
KIT	**24.0**	**(5,660)**
UEDIN	35.7	(8,438)
FBK	38.8	(9,167)

TED : ASR Italian (ASR$_{IT}$)

System	WER	(# Errors)
VECSYS-LIUM	**21.9**	**(5,165)**
MITLL-AFRL	23.0	(5,440)
FBK	23.8	(5618)
KIT	25.4	(5,997)

TED : SLT English-French (SLT$_{EnFr}$)

System	*case sensitive*		*case insensitive*	
	BLEU	TER	BLEU	TER
KIT	**27.45**	57.80	**28.16**	56.87
RWTH	26.94	57.29	27.74	**56.22**
LIUM	26.82	59.03	27.85	57.69
UEDIN	25.50	**57.23**	26.26	56.24
FBK	25.39	59.53	26.11	58.57
LIMSI	25.18	60.70	25.88	59.69
USFD	23.45	59.94	24.14	58.97

TED : SLT English-German (SLT$_{EnDe}$)

System	*case sensitive*		*case insensitive*	
	BLEU	TER	BLEU	TER
KIT	**17.05**	**68.01**	**17.58**	**66.97**
UEDIN	17.00	68.36	17.51	67.30
USFD	14.75	70.15	15.24	69.15
KLE	13.00	71.70	13.64	70.33

TED : SLT German-English (SLT$_{DeEn}$)

System	*case sensitive*		case insensitive	
	BLEU	TER	BLEU	TER
EU-BRIDGE	**19.09**	**63.80**	**19.59**	**62.94**
KIT	18.34	63.91	18.85	62.99
UEDIN	17.67	66.04	18.18	65.12
RWTH	17.24	65.04	17.78	64.07
KLE	9.95	74.05	10.36	72.97

TED : MT English-French (MT$_{EnFr}$)

System	*case sensitive*		*case insensitive*	
	BLEU	TER	BLEU	TER
EU-BRIDGE	**36.99**	45.20	**37.85**	**44.32**
KIT	36.22	**45.18**	36.97	44.37
UEDIN	35.91	45.78	36.64	45.04
RWTH	35.72	44.54	36.46	43.77
MITLL-AFRL	35.48	45.69	36.90	44.49
FBK	34.24	46.75	34.85	46.04
BASELINE	30.55	49.66	31.13	49.00
MIRACL	25.86	54.16	26.97	53.02
SFAX	16.09	62.89	17.33	61.48

TED : MT English-German (MT$_{EnDe}$)

System	*case sensitive* BLEU	*case sensitive* TER	*case insensitive* BLEU	*case insensitive* TER
EU-BRIDGE	**23.25**	**57.27**	**24.06**	**56.15**
KIT	22.66	57.70	23.35	56.66
UEDIN	22.61	58.95	23.14	57.92
NTT-NAIST	22.09	57.60	22.63	56.65
KLE	19.26	61.36	19.75	60.48
BASELINE	18.44	61.89	18.92	61.02

TED : MT German-English (SLT$_{DeEn}$)

System	*case sensitive* BLEU	*case sensitive* TER	*case insensitive* BLEU	*case insensitive* TER
EU-BRIDGE	**25.77**	**54.61**	**26.36**	**53.76**
RWTH	25.04	55.49	25.61	54.65
KIT	24.62	55.62	25.16	54.77
NTT-NAIST	23.77	56.43	24.52	55.49
UEDIN	23.32	57.50	24.06	56.55
FBK	20.52	63.37	21.77	60.66
KLE	19.31	63.88	20.60	61.38
BASELINE	17.50	65.56	18.61	63.08

TED : MT English-Arabic (MT$_{EnAr}$)

System	BLEU	TER
UEDIN	**13.24**	**69.16**
KIT	13.05	71.62
BASELINE	11.12	72.88

TED : MT Arabic-English (MT$_{ArEn}$)

System	*case sensitive* BLEU	*case sensitive* TER	*case insensitive* BLEU	*case insensitive* TER
MITLL-AFRL	**27.52**	**54.54**	**28.41**	**53.44**
UEDIN	25.46	57.07	26.22	56.02
BASELINE	19.88	63.30	20.48	62.31

TED : MT English-Spanish (MT$_{EnEs}$)

System	*case sensitive* BLEU	*case sensitive* TER	*case insensitive* BLEU	*case insensitive* TER
UEDIN	**35.63**	**45.10**	**36.47**	**44.12**
BASELINE	31.26	48.43	31.95	47.48

TED : MT Spanish-English (MT$_{EsEn}$)

System	*case sensitive* BLEU	*case sensitive* TER	*case insensitive* BLEU	*case insensitive* TER
UEDIN	**37.29**	**43.73**	**38.07**	**42.85**
BASELINE	33.31	46.07	33.80	45.38

TED : MT English-Farsi (MT$_{EnFa}$)

System	BLEU	TER
BASELINE	**6.48**	**81.14**

TED : MT Farsi-English (MT$_{FaEn}$)

System	*case sensitive* BLEU	*case sensitive* TER	*case insensitive* BLEU	*case insensitive* TER
MITLL-AFRL	**18.37**	**66.02**	**19.03**	**65.03**
UEDIN	16.94	72.66	17.52	71.66
BASELINE	16.22	72.13	16.72	71.05

TED : MT English-Hebrew (MT$_{EnHe}$)

System	*case sensitive* BLEU	*case sensitive* TER	*case insensitive* BLEU	*case insensitive* TER
BASELINE	**15.69**	**65.62**	**15.69**	**65.62**

TED : MT Hebrew-English (MT$_{HeEn}$)

System	*case sensitive* BLEU	*case sensitive* TER	*case insensitive* BLEU	*case insensitive* TER
UEDIN	**26.58**	**56.99**	**27.14**	**56.25**
BASELINE	23.66	58.66	24.20	57.83

TED : MT English-Polish (MT$_{EnPl}$)

System	*case sensitive* BLEU	*case sensitive* TER	*case insensitive* BLEU	*case insensitive* TER
PJIIT	**16.10**	**74.82**	**16.60**	**73.64**
BASELINE	9.75	82.60	10.16	81.44
LIA	7.79	86.89	10.12	82.31

TED : MT Polish-English (MT$_{PlEn}$)

System	*case sensitive* BLEU	*case sensitive* TER	*case insensitive* BLEU	*case insensitive* TER
PJIIT	**18.33**	**65.60**	**18.96**	**64.59**
BASELINE	13.94	68.75	14.49	67.63

TED : MT English-Portuguese (MT$_{EnPt}$)

System	*case sensitive* BLEU	*case sensitive* TER	*case insensitive* BLEU	*case insensitive* TER
UEDIN	**32.41**	**45.85**	**33.12**	**44.87**
BASELINE	31.32	47.06	31.97	46.19

TED : MT Portuguese-English (MT$_{PtEn}$)

System	*case sensitive* BLEU	*case sensitive* TER	*case insensitive* BLEU	*case insensitive* TER
BASELINE	**35.78**	**43.13**	**36.16**	**42.61**
UEDIN	34.66	46.11	35.28	45.52

TED : MT English-Russian(MT$_{EnRu}$)

System	*case sensitive*		*case insensitive*	
	BLEU	TER	BLEU	TER
BASELINE	**11.21**	**73.15**	**11.21**	**72.24**

TED : MT English-Slovenian(MT$_{EnSl}$)

System	*case sensitive*		*case insensitive*	
	BLEU	TER	BLEU	TER
LIA	**10.36**	**71.81**	**12.69**	**67.80**
BASELINE	8.53	73.75	8.87	72.76

TED : MT English-Turkish(MT$_{EnTr}$)

System	*case sensitive*		*case insensitive*	
	BLEU	TER	BLEU	TER
BASELINE	**6.97**	**79.93**	**7.36**	**78.65**
UMONTREAL	4.76	80.67	5.51	79.28

TED : MT English-Chinese(MT$_{EnZh}$)

System	*character-based*	
	BLEU	TER
USTC	21.64	65.71
KIT	18.31	66.43
HKUST	16.41	74.35
BASELINE	15.56	80.48
UMONTREAL	7.40	81.89

TED : MT Russian-English (MT$_{RuEn}$)

System	*case sensitive*		*case insensitive*	
	BLEU	TER	BLEU	TER
MITLL-AFRL	**19.30**	**63.95**	**20.22**	**62.64**
BASELINE	15.48	69.93	15.95	68.91

TED : MT Slovenian-English (MT$_{SlEn}$)

System	*case sensitive*		*case insensitive*	
	BLEU	TER	BLEU	TER
BASELINE	**13.69**	**70.79**	**14.07**	**69.83**

TED : MT Turkish-English (MT$_{TrEn}$)

System	*case sensitive*		*case insensitive*	
	BLEU	TER	BLEU	TER
BASELINE	**12.52**	**76.96**	**13.10**	**75.77**

TED : MT Chinese-English (MT$_{ZhEn}$)

System	*case sensitive*		*case insensitive*	
	BLEU	TER	BLEU	TER
USTC	**15.65**	**69.65**	**16.35**	**68.62**
NICT	14.05	71.68	14.88	70.42
MITLL-AFRL	12.83	74.74	13.51	73.58
BASELINE	11.22	72.43	11.79	71.37
HKUST	9.64	76.67	10.83	74.16

A.2. Progress Test Set (*tst2013*)

· All the sentence IDs in the IWSLT 2013 test set were used to calculate the automatic scores for each run submission.
· ASR and MT systems are ordered according to the *WER* and *BLEU* metrics, respectively.
· For each task, the best score of each metric is marked with **boldface**.
· All automatic evaluation metric scores are given as percent figures (%).

TED: ASR English tst2013

System	IWSLT 2013 WER	IWSLT 2013 (# Errors)	IWSLT 2014 WER	IWSLT 2014 (# Errors)
NICT	**13.5**	**(5,734)**	**10.6**	**(4,518)**
MITLL-AFRL	15.9	(6,788)	13.7	(5,856)
KIT	14.4	(6,115)	14.2	(6,044)
FBK	23.2	(9,899)	14.7	(6,247)
LIUM	—		16.0	(6,818)
UEDIN	22.1	(9,413)	16.3	(6,963)
IOIT	27.2	(11,578)	24.0	(10,206)

TED: ASR German tst2013

System	IWSLT 2013 WER	IWSLT 2013 (# Errors)	IWSLT 2014 WER	IWSLT 2014 (# Errors)
KIT	**25.7**	**(4,932)**	**25.4**	**(5,885)**
UEDIN	37.8	(7,250)	35.0	(6,720)
FBK	37.5	(7,199)	37.8	(7,261)

TED : MT English-French test 2013(MT$_{EnFr}$)

System	*case sensitive* BLEU	*case sensitive* TER	*case insensitive* BLEU	*case insensitive* TER
EU-BRIDGE	**40.50**	43.27	**41.65**	42.06
KIT	40.12	43.09	41.11	42.04
RWTH	39.72	**42.58**	40.73	**41.52**
UEDIN	39.59	43.80	40.45	42.78
MITLL-AFRL	39.08	44.05	40.59	42.73
FBK	38.20	44.83	38.99	43.88
BASELINE	33.20	48.91	33.81	48.07
MIRACL	29.63	51.96	30.91	50.65

TED : MT English-German test 2013 (MT$_{EnDe}$)

System	*case sensitive* BLEU	*case sensitive* TER	*case insensitive* BLEU	*case insensitive* TER
EU-BRIDGE	**26.22**	**53.62**	**27.30**	**52.34**
KIT	26.03	53.83	26.77	52.81
NTT-NAIST	25.80	53.86	26.55	52.75
UEDIN	25.33	55.12	26.13	53.93
KLE	21.69	58.27	22.25	57.32
BASELINE	20.96	58.48	21.52	57.58

TED : MT German-English test 2013 (MT$_{DeEn}$)

System	*case sensitive* BLEU	*case sensitive* TER	*case insensitive* BLEU	*case insensitive* TER
EU-BRIDGE	**28.77**	**50.52**	**29.29**	**49.63**
KIT	27.98	50.92	28.55	50.04
NTT-NAIST	27.81	51.62	28.32	50.82
UEDIN	27.60	52.43	28.26	51.44
RWTH	27.59	51.33	28.08	50.41
FBK	25.45	55.80	26.07	54.88
KLE	23.59	57.38	24.18	56.47
BASELINE	20.26	60.33	20.89	59.48

TED : MT English-Arabic test 2013(MT$_{EnAr}$)

System	BLEU	TER
UEDIN	**14.20**	**65.97**
KIT	14.15	68.29
BASELINE	12.68	68.94

TED : MT Arabic-English test 2013 (MT$_{ArEn}$)

System	*case sensitive* BLEU	*case sensitive* TER	*case insensitive* BLEU	*case insensitive* TER
MITLL-AFRL	**31.48**	**49.88**	**32.41**	**48.76**
UEDIN	29.06	53.02	29.74	52.03
BASELINE	21.63	60.32	22.46	59.12

TED : MT English-Spanish test 2013 (MT$_{EnEs}$)

System	*case sensitive* BLEU	*case sensitive* TER	*case insensitive* BLEU	*case insensitive* TER
UEDIN	**34.74**	**45.75**	**35.42**	**44.78**
BASELINE	30.63	49.39	31.14	48.57

TED : MT Spanish-English test 2013(MT$_{EsEn}$)

System	*case sensitive* BLEU	*case sensitive* TER	*case insensitive* BLEU	*case insensitive* TER
UEDIN	**39.13**	**41.37**	**39.75**	**40.60**
BASELINE	34.18	44.63	34.70	44.00

TED : MT English-Farsi test 2013 (MT$_{EnFa}$)

System	BLEU	TER
BASELINE	7.05	78.90

TED : MT Farsi-English test 2013 (MT$_{FaEn}$)

System	*case sensitive*		*case insensitive*	
	BLEU	TER	BLEU	TER
MITLL-AFRL	**19.47**	**63.27**	**20.11**	**62.27**
UEDIN	16.51	82.50	16.87	81.58
BASELINE	14.04	83.01	14.44	82.09

TED : MT English-Hebrew test 2013(MT$_{EnHe}$)

System	*case sensitive*		*case insensitive*	
	BLEU	TER	BLEU	TER
BASELINE	**15.92**	**64.16**	**15.92**	**64.16**

TED : MT Hebrew-English test2013 (MT$_{HeEn}$)

System	*case sensitive*		*case insensitive*	
	BLEU	TER	BLEU	TER
UEDIN	**29.70**	**52.40**	**30.51**	**51.35**
BASELINE	25.97	55.40	26.74	54.23

TED : MT English-Polish test2013 (MT$_{EnPl}$)

System	*case sensitive*		*case insensitive*	
	BLEU	TER	BLEU	TER
PJIIT	**25.92**	**61.04**	**26.62**	**59.94**
BASELINE	11.12	75.95	11.67	74.78

TED : MT Polish-English test2013 (MT$_{PlEn}$)

System	*case sensitive*		*case insensitive*	
	BLEU	TER	BLEU	TER
PJIIT	**27.99**	**58.01**	**28.61**	**57.10**
BASELINE	17.25	66.44	17.75	65.44

TED : MT English-Portuguese test 2013(MT$_{EnPt}$)

System	*case sensitive*		*case insensitive*	
	BLEU	TER	BLEU	TER
BASELINE	31.38	46.42	31.89	45.66
UEDIN	33.20	44.90	33.93	43.90

TED : MT Portuguese-English test 2013 (MT$_{PtEn}$)

System	*case sensitive*		*case insensitive*	
	BLEU	TER	BLEU	TER
BASELINE	37.85	40.87	38.26	40.35
UEDIN	37.34	42.91	37.80	42.30

TED : MT English-Russian test 2013(MT$_{EnRu}$)

System	*case sensitive*		*case insensitive*	
	BLEU	TER	BLEU	TER
BASELINE	**14.01**	**70.47**	**14.01**	**69.44**

TED : MT Russian-English test 2012 (MT$_{RuEn}$)

System	*case sensitive*		*case insensitive*	
	BLEU	TER	BLEU	TER
MITLL-AFRL	**24.30**	**57.59**	**25.39**	**56.25**
BASELINE	19.82	63.56	20.40	62.46

TED : MT English-Slovenian test 2013 (MT$_{EnSl}$)

System	*case sensitive*		*case insensitive*	
	BLEU	TER	BLEU	TER
BASELINE	**9.63**	**73.32**	**9.97**	**72.34**

TED : MT Slovenian-English test2013 (MT$_{SlEn}$)

System	*case sensitive*		*case insensitive*	
	BLEU	TER	BLEU	TER
BASELINE	**14.64**	**68.68**	**15.19**	**67.63**

TED : MT English-Turkish test 2013 (MT$_{EnTr}$)

System	*case sensitive*		*case insensitive*	
	BLEU	TER	BLEU	TER
BASELINE	**6.85**	**80.40**	**7.21**	**79.08**
UMONTREAL	4.06	83.97	4.77	82.50

TED : MT Turkish-English test 2013 (MT$_{TrEn}$)

System	*case sensitive*		*case insensitive*	
	BLEU	TER	BLEU	TER
BASELINE	**13.30**	**75.17**	**13.95**	**74.00**

TED : MT English-Chinese test2013 (MT$_{EnZh}$)

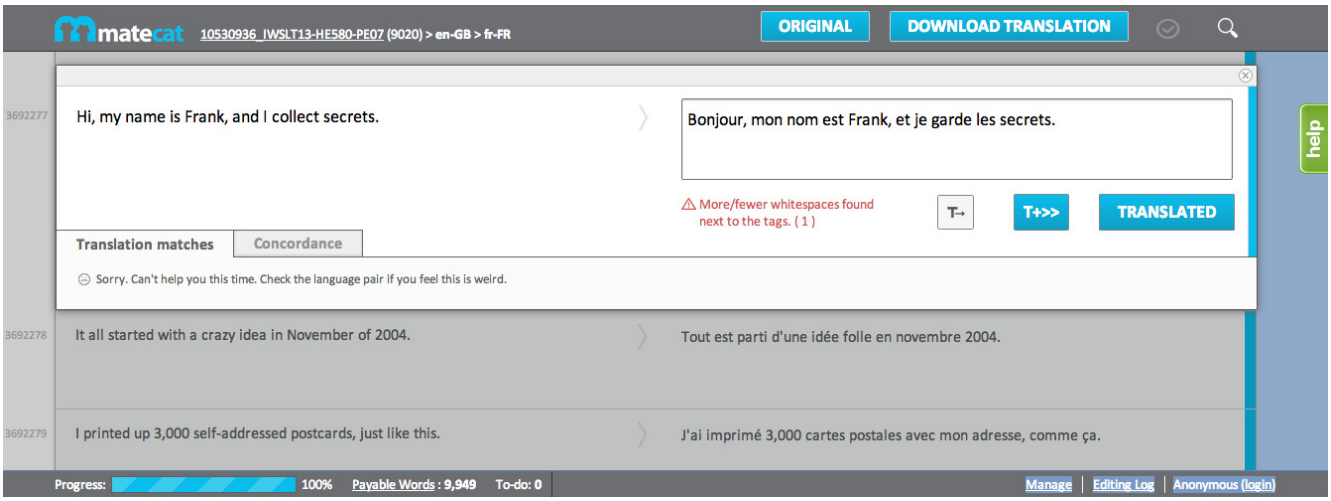
System	*character-based*	
	BLEU	TER
USTC	**22.49**	**63.74**
KIT	21.01	63.12
HKUST	18.81	70.94
BASELINE	18.23	76.15
UMONTREAL	7.93	80.47

TED : MT Chinese-English test 2013(MT$_{ZhEn}$)

System	*case sensitive*		*case insensitive*	
	BLEU	TER	BLEU	TER
USTC	**18.12**	**66.28**	**18.85**	**65.23**
NICT	16.57	67.96	17.36	66.76
MITLL-AFRL	15.59	70.89	16.32	69.68
BASELINE	13.40	68.85	14.00	67.90
HKUST	11.89	72.33	13.08	70.10

Appendix B. Human Evaluation

Interface used for the bilingual post-editing task



Post-editing instructions given to professional translators

In this task you are presented with automatic translations of TED Talks captions.

You are asked to post-edit the given automatic translation by applying the minimal edits required to transform the system output into a fluent sentence with the same meaning as the source sentence.

While post-editing, remember that the post-edited sentence is to be intended as a transcription of spoken language. Note also that the focus is the correctness of the single sentence within the given context, NOT the consistency of a group of sentences. Hence, surrounding segments should be used to understand the context but NOT to enforce consistency on the use of terms. In particular, different but correct translations of terms across segments should not be corrected.

Examples:

Source: This next one takes a little explanation before I share it with you.
Automatic translation: ...avant que je partage avec vous.
Post-editing 1: ...avant de le partager avec vous.
Post-editing 2: ...avant que je le partage avec vous. (preferred - minimal editing and acceptable in spoken language)

Source: And the table form is important.
Automatic translation: Et la forme de la table est importante.
Post-editing 1: La forme de la table est également importante.
Post-editing 2: Et la forme de la table est importante. (preferred - no editing - slightly less fluent but better fitting the source speech transcription)

Source: Everyone who knew me before 9/11 believes...
Automatic translation: ...avant le 11/9...
Post-editing 1: ...avant le 11 septembre...
Post-editing 2: ...avant le 11/9... (preferred - no editing - better fitting the source)