Findings of the 2015 Workshop on Statistical Machine Translation

Ondřej Bojar
Charles Univ. in Prague

Rajen Chatterjee
FBK

Christian Federmann
Microsoft Research

Barry Haddow
Univ. of Edinburgh

Matthias Huck
Univ. of Edinburgh

Chris Hokamp
Dublin City Univ.

Philipp Koehn
JHU / Edinburgh

Varvara Logacheva
Univ. of Sheffield

Christof Monz
Univ. of Amsterdam

Matteo Negri
FBK

Matt Post
Johns Hopkins Univ.

Carolina Scarton
Univ. of Sheffield

Lucia Specia
Univ. of Sheffield

Marco Turchi
FBK

Abstract

This paper presents the results of the WMT15 shared tasks, which included a standard news translation task, a metrics task, a tuning task, a task for run-time estimation of machine translation quality, and an automatic post-editing task. This year, 68 machine translation systems from 24 institutions were submitted to the ten translation directions in the standard translation task. An additional 7 anonymized systems were included, and were then evaluated both automatically and manually. The quality estimation task had three subtasks, with a total of 10 teams, submitting 34 entries. The pilot automatic post-editing task had a total of 4 teams, submitting 7 entries.

1 Introduction

We present the results of the shared tasks of the Workshop on Statistical Machine Translation (WMT) held at EMNLP 2015. This workshop builds on eight previous WMT workshops (Koehn and Monz, 2006; Callison-Burch et al., 2007, 2008, 2009, 2010, 2011, 2012; Bojar et al., 2013, 2014). This year we conducted five official tasks: a translation task, a quality estimation task, a metrics task, a tuning task[1], and a automatic post-editing task.

In the translation task (§2), participants were asked to translate a shared test set, optionally restricting themselves to the provided training data. We held ten translation tasks this year, between English and each of Czech, French, German, Finnish, and Russian. The Finnish translation tasks were new this year, providing a lesser resourced data condition on a challenging language pair. The system outputs for each task were evaluated both automatically and manually.

The human evaluation (§3) involves asking human judges to rank sentences output by anonymized systems. We obtained large numbers of rankings from researchers who contributed evaluations proportional to the number of tasks they entered. We made data collection more efficient and used TrueSkill as ranking method.

The quality estimation task (§4) this year included three subtasks: sentence-level prediction of post-editing effort scores, word-level prediction of good/bad labels, and document-level prediction of Meteor scores. Datasets were released with English→Spanish news translations for sentence and word level, English↔German news translations for document level.

The first round of the automatic post-editing task (§5) examined automatic methods for correcting errors produced by an unknown machine translation system. Participants were provided with training triples containing source, target and human post-editions, and were asked to return automatic post-editions for unseen (source, target) pairs. This year we focused on correcting English→Spanish news translations.

The primary objectives of WMT are to evaluate the state of the art in machine translation, to disseminate common test sets and public training data with published performance numbers, and to refine evaluation and estimation methodologies for machine translation. As before, all of the data, translations, and collected human judgments are publicly available.[2] We hope these datasets serve as a valuable resource for research into statistical

[1] The metrics and tuning tasks are reported in separate papers (Stanojević et al., 2015a,b).

[2] http://statmt.org/wmt15/results.html

Proceedings of the Tenth Workshop on Statistical Machine Translation, pages 1–46,
Lisboa, Portugal, 17-18 September 2015.

machine translation and automatic evaluation or prediction of translation quality.

2 Overview of the Translation Task

The recurring task of the workshop examines translation between English and other languages. As in the previous years, the other languages include German, French, Czech and Russian.

Finnish replaced Hindi as the special language this year. Finnish is a lesser resourced language compared to the other languages and has challenging morphological properties. Finnish represents also a different language family that we had not tackled since we included Hungarian in 2008 and 2009 (Callison-Burch et al., 2008, 2009).

We created a test set for each language pair by translating newspaper articles and provided training data, except for French, where the test set was drawn from user-generated comments on the news articles.

2.1 Test data

The test data for this year's task was selected from online sources, as before. We took about 1500 English sentences and translated them into the other 5 languages, and then additional 1500 sentences from each of the other languages and translated them into English. This gave us test sets of about 3000 sentences for our English-X language pairs, which have been either written originally written in English and translated into X, or vice versa.

For the French-English discussion forum test set, we collected 38 discussion threads each from the Guardian for English and from Le Monde for French. See Figure 1 for an example.

The composition of the test documents is shown in Table 1.

The stories were translated by the professional translation agency Capita, funded by the EU Framework Programme 7 project MosesCore, and by Yandex, a Russian search engine company.[3] All of the translations were done directly, and not via an intermediate language.

2.2 Training data

As in past years we provided parallel corpora to train translation models, monolingual corpora to train language models, and development sets to tune system parameters. Some training corpora were identical from last year (Europarl[4], United Nations, French-English 10^9 corpus, CzEng, Common Crawl, Russian-English parallel data provided by Yandex, Russian-English Wikipedia Headlines provided by CMU), some were updated (News Commentary, monolingual data), and new corpora was added (Finnish Europarl), Finnish-English Wikipedia Headline corpus).

Some statistics about the training materials are given in Figure 2.

2.3 Submitted systems

We received 68 submissions from 24 institutions. The participating institutions and their entry names are listed in Table 2; each system did not necessarily appear in all translation tasks. We also included 1 commercial off-the-shelf MT system and 6 online statistical MT systems, which we anonymized.

For presentation of the results, systems are treated as either *constrained* or *unconstrained*, depending on whether their models were trained only on the provided data. Since we do not know how they were built, these online and commercial systems are treated as unconstrained during the automatic and human evaluations.

3 Human Evaluation

Following what we had done for previous workshops, we again conduct a human evaluation campaign to assess translation quality and determine the final ranking of candidate systems. This section describes how we prepared the evaluation data, collected human assessments, and computed the final results.

This year's evaluation campaign differed from last year in several ways:

- In previous years each ranking task compared five different candidate systems which were selected without any pruning or redundancy cleanup. This had resulted in a noticeable amount of near-identical ranking candidates in WMT14, making the evaluation process unnecessarily tedious as annotators ran into a fair amount of ranking tasks containing very similar segments which are hard to inspect. For WMT15, we perform redundancy cleanup as an initial preprocessing step and

[3] http://www.yandex.com/

[4] As of Fall 2011, the proceedings of the European Parliament are no longer translated into all official languages.

This is perfectly illustrated by the UKIP numbties banning people with HIV.
You mean Nigel Farage saying the NHS should not be used to pay for people coming to the UK as health tourists, and saying yes when the interviewer specifically asked if, with the aforementioned in mind, people with HIV were included in not being welcome.
You raise a straw man and then knock it down with thinly veiled homophobia.
Every time I or my family need to use the NHS we have to queue up behind bigots with a sense of entitlement and chronic hypochondria.
I think the straw man is yours.
Health tourism as defined by the right wing loonies is virtually none existent.
I think it's called democracy.
So no one would be affected by UKIP's policies against health tourism so no problem.
Only in UKIP La La Land could Carswell be described as revolutionary.
Quoting the bollox The Daily Muck spew out is not evidence.
Ah, shoot the messenger.
The Mail didn't write the report, it merely commented on it.
Whoever controls most of the media in this country should undead be shot for spouting populist propaganda as fact.
I don't think you know what a straw man is.
You also don't know anything about my personal circumstances or identity so I would be very careful about trying to eradicate a debate with accusations of homophobia.
Farage's comment came as quite a shock, but only because it is so rarely addressed.
He did not express any homophobic beliefs whatsoever.
You will just have to find a way of getting over it.
I'm not entirely sure what you're trying to say, but my guess is that you dislike the media reporting things you disagree with.
It is so rarely addressed because unlike Fararge and his Thatcherite loony disciples who think aids and floods are a signal from the divine and not a reflection on their own ignorance in understanding the complexities of humanity as something to celebrate,then no.

Figure 1: Example news discussion thread used in the French–English translation task.

Language	Sources (Number of Documents)
Czech	aktuálně.cz (4), blesk.cz (1), blisty.cz (1), ctk.cz (1), deník.cz (1), e15.cz (1), iDNES.cz (19), ihned.cz (3), lidovky.cz (6), Novinky.cz (2), tyden.cz (1).
English	ABC News (4), BBC (6), CBS News (1), Daily Mail (1), Euronews (1), Financial Times (1), Fox News (2), Globe and Mail (1), Independent (1), Los Angeles Times (1), News.com Australia (9), Novinite (2), Reuters (2), Sydney Morning Herald (1), stv.tv (1), Telegraph (8), The Local (1), The Nation (1), UPI (1), Washington Post (3).
German	Abendzeitung Nürnberg (1), Aachener Nachrichten (1), Der Standard (2), Deutsche Welle (1), Frankfurter Neue Presse (1), Frankfurter Rundschau (1), Generalanzeiger Bonn (2), Göttinger Tageblatt (1), Haller Kreisblatt (1), Hellweger Anzeiger (1), Junge Welt (1), Kreisanzeiger (1), Mainpost (1), Merkur (3), Mittelbayerische Nachrichten (2), Morgenpost (1), Mitteldeutsche Zeitung (1), Neue Presse Coburg (1), Nürtinger Zeitung (1), OE24 (1), Kölnische Rundschau (1), Tagesspiegel (1), Volksfreund (1), Volksstimme (1), Wiener Zeitung (1), Westfälische Nachrichten (2).
Finnish	Aamulehti (2), Etelä-Saimaa (1), Etelä-Suomen Sanomat (3), Helsingin Sanomat (13), Ilkka (7), Ilta-Sanomat (18), Kaleva (4), Karjalainen (2), Kouvolan Sanomat (1), Lapin Kansa (3), Maaseudun Tulevaisuus (1).
Russian	168.ru (1), aif (6), altapress.ru (1), argumenti.ru (8), BBC Russian (1), dp.ru (2), gazeta.ru (4), interfax (2), Kommersant (12), lenta.ru (8), lgng (3), mk (5), novinite.ru (1), rbc.ru (1), rg.ru (2), rusplit.ru (1), Sport Express (6), vesti.ru (10).

Table 1: Composition of the test set. For more details see the XML test files. The `docid` tag gives the source and the date for each document in the test set, and the `origlang` tag indicates the original source language.

Europarl Parallel Corpus

	French ↔ English		German ↔ English		Czech ↔ English		Finnish ↔ English	
Sentences	2,007,723		1,920,209		646,605		1,926,114	
Words	60,125,563	55,642,101	50,486,398	53,008,851	14,946,399	17,376,433	37,814,266	52,723,296
Distinct words	140,915	118,404	381,583	115,966	172,461	63,039	693,963	115,896

News Commentary Parallel Corpus

	French ↔ English		German ↔ English		Czech ↔ English		Russian ↔ English	
Sentences	200,239		216,190		152,763		174,253	
Words	6,270,748	5,161,906	5,513,985	5,499,625	3,435,458	3,759,874	4,394,974	4,625,898
Distinct words	75,462	71,767	157,682	74,341	142,943	58,817	172,021	67,402

Common Crawl Parallel Corpus

	French ↔ English		German ↔ English		Czech ↔ English		Russian ↔ English	
Sentences	3,244,152		2,399,123		161,838		878,386	
Words	91,328,790	81,096,306	54,575,405	58,870,638	3,529,783	3,927,378	21,018,793	21,535,122
Distinct words	889,291	859,017	1,640,835	823,480	210,170	128,212	764,203	432,062

United Nations Parallel Corpus

	French ↔ English	
Sentences	12,886,831	
Words	411,916,781	360,341,450
Distinct words	565,553	666,077

Yandex 1M Parallel Corpus

	Russian ↔ English	
Sentences	1,000,000	
Words	24,121,459	26,107,293
Distinct words	701,809	387,646

10^9 Word Parallel Corpus

	French ↔ English	
Sentences	22,520,400	
Words	811,203,407	668,412,817
Distinct words	2,738,882	2,861,836

CzEng Parallel Corpus

	Czech ↔ English	
Sentences	14,833,358	
Words	200,658,857	228,040,794
Distinct words	1,389,803	920,824

Wiki Headlines Parallel Corpus

	Russian ↔ English		Finnish ↔ English	
Sentences	514,859		153,728	
Words	1,191,474	1,230,644	269,429	354,362
Distinct words	282,989	251,328	127,576	96,732

Europarl Language Model Data

	English	French	German	Czech	Finnish
Sentence	2,218,201	2,190,579	2,176,537	668,595	2,120,739
Words	59,848,044	63,439,791	53,534,167	14,946,399	39,511,068
Distinct words	123,059	145,496	394,781	172,461	711,868

News Language Model Data

	English	French	German	Czech	Russian	Finnish
Sentence	118,337,431	42,110,011	135,693,607	45,149,206	45,835,812	1,378,582
Words	2,744,428,620	1,025,132,098	2,427,581,519	745,645,366	823,284,188	16,501,511
Distinct words	4,895,080	2,352,451	13,727,336	3,513,784	3,885,756	925,201

Test Set

	French ↔ English		German ↔ English		Czech ↔ English		Russian ↔ English		Finnish ↔ English	
Sentences	1500		2169		2656		2818		1370	
Words	29,858	27,173	44,081	46,828	46,005	54,055	55,655	65,744	19,840	27,811
Distinct words	5,798	5,148	9,710	7,483	13,013	7,757	15,795	8,695	8,553	5,279

Figure 2: Statistics for the training and test sets used in the translation task. The number of words and the number of distinct words (case-insensitive) is based on the provided tokenizer.

ID	Institution
AALTO	Aalto University (Grönroos et al., 2015)
ABUMATRAN	Abu-MaTran (Rubino et al., 2015)
AFRL-MIT-*	Air Force Research Laboratory / MIT Lincoln Lab (Gwinnup et al., 2015)
CHALMERS	Chalmers University of Technology (Kolachina and Ranta, 2015)
CIMS	University of Stuttgart and Munich (Cap et al., 2015)
CMU	Carnegie Mellon University
CU-CHIMERA	Charles University (Bojar and Tamchyna, 2015)
CU-TECTO	Charles University (Dušek et al., 2015)
DFKI	Deutsches Forschungszentrum für Künstliche Intelligenz (Avramidis et al., 2015)
ILLINOIS	University of Illinois (Schwartz et al., 2015)
IMS	University of Stuttgart (Quernheim, 2015)
KIT	Karsruhe Institut of Technology (Cho et al., 2015)
KIT-LIMSI	Karsruhe Institut of Technology / LIMSI (Ha et al., 2015)
LIMSI	LIMSI (Marie et al., 2015)
MACAU	University of Macau
MONTREAL	University of Montreal (Jean et al., 2015)
PROMT	ProMT
RWTH	RWTH Aachen (Peter et al., 2015)
SHEFF*	University of Sheffield (Steele et al., 2015)
UDS-SANT	University of Saarland (Pal et al., 2015a)
UEDIN-JHU	University of Edinburgh / Johns Hopkins University (Haddow et al., 2015)
UEDIN-SYNTAX	University of Edinburgh (Williams et al., 2015)
USAAR-GACHA	University of Saarland, Liling Tan
UU	Uppsala University (Tiedemann et al., 2015)
COMMERCIAL-1	Commercial machine translation system
ONLINE-[A,B,C,E,F,G]	Six online statistical machine translation systems

Table 2: Participants in the shared translation task. Not all teams participated in all language pairs. The translations from the commercial and online systems were not submitted by their respective companies but were obtained by us, and are therefore anonymized in a fashion consistent with previous years of the workshop.

create *multi-system translations*. As a consequence, we get ranking tasks with varying numbers of candidate systems. To avoid overloading the annotators we still allow a maximum of five candidates per ranking task. If we have more multi-system translations, we choose randomly.

A brief example should illustrate this more clearly: say we have the following two candidate systems:

```
sysA="This, is 'Magic'"
sysX="this is  magic"
```

After lowercasing, removal of punctuation and whitespace normalization, which are our criteria for identifying *near-identical* outputs, both would be collapsed into a single multi-system:

```
sysA+sysX="This, is 'Magic'"
```

The first representative of a group of near-identical outputs is used as a proxy representing all candidates in the group throughout the evaluation.

While there is a good chance that users would have used some of the stripped information, e.g., case to differentiate between the two systems relative to each other, the collapsed system's comparison result against the other candidates should be a good approximation of how human annotators would have ranked them individually. We get a near 2x increase in the number of pairwise comparisons, so the general approach seems helpful.

- After dropping external, crowd-sourced translation assessment in WMT14 we ended up with approximately seventy-five percent less raw comparison data. Still, we were able to compute good confidence intervals on the clusters based on our improved ranking approach.

  This year, due to the aforementioned cleanup, annotators spent their time more efficiently, resulting in an increased number of final ranking results. We collected a total of 542,732 individual "A > B" judgments this year, nearly double the amount of data compared to WMT14.

- Last year we compared three different models of producing the final system rankings: Expected Wins (used in WMT13), Hopkins and May (HM) and TrueSkill (TS). Overall, we found the TrueSkill method to work best which is why we decided to use it as our only approach in WMT15.

We keep using clusters in our final system rankings, providing a *partial ordering* (clustering) of all evaluated candidate systems. Semantics remain unchanged to previous years: systems in the same cluster could not be meaningfully distinguished and hence are considered to be of equal quality.

3.1 Evaluation campaign overview

WMT15 featured the largest evaluation campaign to date. Similar to last year, we decided to collect *researcher-based judgments* only. A total of 137 individual annotator accounts have been actively involved. Users came from 24 different research groups and contributed judgments on 9,669 HITs.

Overall, these correspond to 29,007 individual ranking tasks (plus some more from incomplete HITs), each of which would have spawned exactly 10 individual "A > B" judgments last year, so we expected at least >290,070 binary data points. Due to our redundancy cleanup, we are able to get a lot more, namely 542,732. We report our inter/intra-annotator agreement scores based on the actual work done (*otherwise, we'd artificially boost scores based on inferred rankings*) and use the full set of data to compute clusters (*where the inferred rankings contribute meaningful data*).

Human annotation effort was exceptional and we are grateful to all participating individuals and teams. We believe that human rankings provide the best decision basis for machine translation evaluation and it is great to see contributions on this large a scale. In total, our human annotators spent 32 days and 20 hours working in Appraise.

The average annotation time per HIT amounts to 4 minutes 53 seconds. Several annotators passed the mark of 100 HITs annotated, some worked for more than 24 hours. We don't take this enormous amount of effort for granted and will make sure to improve the evaluation platform and overall process for upcoming workshops.

3.2 Data collection

The system ranking is produced from a large set of pairwise judgments on the translation quality of

candidate systems. Annotations are collected in an evaluation campaign that enlists participants in the shared task to help. Each team is asked to contribute one hundred "Human Intelligence Tasks" (HITs) per primary system submitted.

Each HIT consists of three so-called *ranking tasks*. In a ranking task, an annotator is presented with a source segment, a human reference translation, and the outputs of *up to five anonymized candidate systems*, randomly selected from the set of participating systems, and displayed in random order. This year, we perform redundancy cleanup as an initial preprocessing step and create *multi-system translations*. As a consequence, we get ranking tasks with varying numbers of candidate outputs.

There are two main benefits to this approach:

- Annotators are more efficient as they don't have to deal with near-identical translations which are notoriously hard to differentiate; and
- Potentially, we get higher quality annotations as near-identical systems will be assigned the same "A > B" ranks, improving consistency.

As in previous years, the evaluation campaign is conducted using Appraise[5] (Federmann, 2012), an open-source tool built using Python's Django framework. At the top of each HIT, the following instructions are provided:

> *You are shown a source sentence followed by several candidate translations. Your task is to rank the translations from best to worst (ties are allowed).*

Annotators can decide to skip a ranking task but are instructed to do this only as a last resort, e.g., if the translation candidates shown on screen are clearly misformatted or contain data issues (wrong language or similar problems). Only a small number of ranking tasks has been skipped in WMT15. A screenshot of the Appraise ranking interface is shown in Figure 3.

Annotators are asked to rank the outputs from 1 (best) to 5 (worst), with ties permitted. Note that a *lower* rank is better. The joint rankings provided by a ranking task are then reduced to the fully expanded set of *pairwise rankings* produced by considering all $\binom{n}{2} \leq 10$ combinations of all $n \leq 5$ outputs in the respective ranking task.

[5] https://github.com/cfedermann/Appraise

For example, consider the following annotation provided among outputs A, B, F, H, and J:

	1	2	3	4	5
F				•	
A				•	
B		•			
J					•
H			•		

As the number of outputs n depends on the number of corresponding *multi-system translations* in the original data, we get varying numbers of resulting binary judgments. Assuming that outputs A and F from above are actually *near-identical*, the annotator this year would see a shorter ranking task:

	1	2	3	4	5
AF				•	
B		•			
J					•
H			•		

Note that AF is a *multi-system translation* covering two candidate systems.

Both examples would be reduced to the following set of pairwise judgments:

$$A > B, A = F, A > H, A < J$$
$$B < F, B < H, B < J$$
$$F > H, F < J$$
$$H < J$$

Here, $A > B$ should be read is "A is ranked higher than (worse than) B". Note that by this procedure, the absolute value of ranks and the magnitude of their differences are discarded. Our WMT15 approach including redundancy cleanup allows to obtain these judgments at a lower cognitive cost for the annotators. This partially explains why we were able to collect more results this year.

For WMT13, nearly a million pairwise annotations were collected from both researchers and paid workers on Amazon's Mechanical Turk, in a roughly 1:2 ratio. Last year, we collected data from researchers only, an ability that was enabled by the use of TrueSkill for producing the partial ranking for each task (§3.4). This year, based on our redundancy cleanup we were able to nearly double the amount of annotations, collecting 542,732. See Table 3 for more details.

3.3 Annotator agreement

Each year we calculate annotator agreement scores for the human evaluation as a measure of

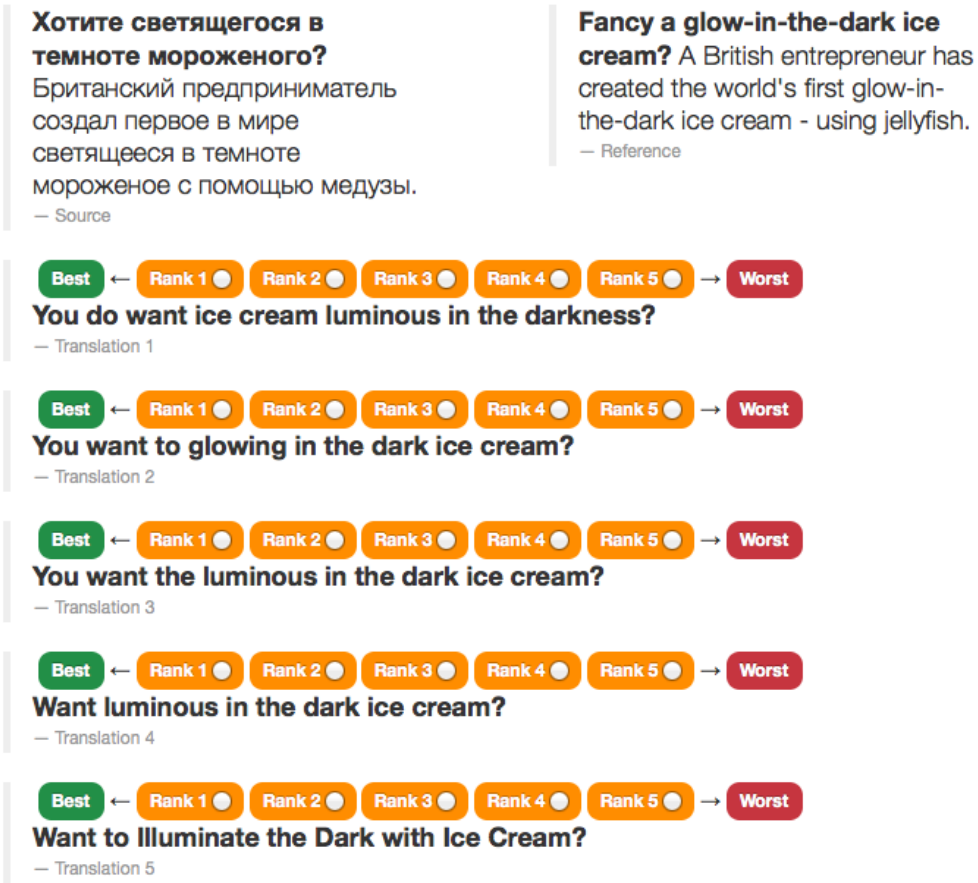

Figure 3: Screenshot of the Appraise interface used in the human evaluation campaign. The annotator is presented with a source segment, a reference translation, and up to five outputs from competing systems (anonymized and displayed in random order), and is asked to rank these according to their translation quality, with ties allowed.

the reliability of the rankings. We measured pairwise agreement among annotators using Cohen's kappa coefficient (κ) (Cohen, 1960). If $P(A)$ be the proportion of times that the annotators agree, and $P(E)$ is the proportion of time that they would agree by chance, then Cohen's kappa is:

$$\kappa = \frac{P(A) - P(E)}{1 - P(E)}$$

Note that κ is basically a normalized version of $P(A)$, one which takes into account how meaningful it is for annotators to agree with each other by incorporating $P(E)$. The values for κ range from 0 to 1, with zero indicating no agreement and 1 perfect agreement.

We calculate $P(A)$ by examining all pairs of outputs[6] which had been judged by two or more judges, and calculating the proportion of time that they agreed that $A < B$, $A = B$, or $A > B$. In other words, $P(A)$ is the empirical, observed rate at which annotators agree, in the context of pairwise comparisons.

As for $P(E)$, it captures the probability that two annotators would agree randomly. Therefore:

$$P(E) = P(A<B)^2 + P(A=B)^2 + P(A>B)^2$$

Note that each of the three probabilities in $P(E)$'s definition are squared to reflect the fact that we are considering the chance that *two* annotators would agree by chance. Each of these probabilities is computed empirically, by observing how often annotators actually rank two systems as being tied.

Table 4 shows final κ values for inter-annotator agreement for WMT11–WMT15 while Table 5 details intra-annotator agreement scores, including the division of researchers (WMT13_r) and MTurk (WMT13_m) data. The exact interpretation of the kappa coefficient is difficult, but according to Landis and Koch (1977), 0–0.2 is *slight*, 0.2–0.4 is *fair*, 0.4–0.6 is *moderate*, 0.6–0.8 is *substantial*, and 0.8–1.0 is *almost perfect*.

[6] regardless if they correspond to an individual system or to a set of systems ("multi-system") producing nearly identical translations

Language Pair	Systems	Rankings	Average
Czech→English	17	85,877	5,051.6
English→Czech	16	136,869	8,554.3
German→English	14	40,535	2,895.4
English→German	17	55,123	3,242.5
French→English	8	29,770	3,721.3
English→French	8	34,512	4,314.0
Russian→English	14	46,193	3,299.5
English→Russian	11	49,582	4,507.5
Finnish→English	15	31,577	2,105.1
English→Finnish	11	32,694	2,972.2
Totals WMT15	131	542,732	4,143.0
WMT14	110	328,830	2,989.3
WMT13	148	942,840	6,370.5
WMT12	103	101,969	999.6
WMT11	133	63,045	474.0

Table 3: Amount of data collected in the WMT15 manual evaluation campagin. The final four rows report summary information from previous editions of the workshop. Note how many rankings we get for Czech language pairs. These include systems from the tuning shared task. Finnish, as a new language, sees a shortage of rankings for Finnnish→English Interest in French seems to have lowered this year with only seven systems. Overall, we see a nice increase in pairwise rankings, especially considering that we have dropped crowd-source annotation and are instead relying on researchers' judgments exclusively.

The inter-annotator agreement rates improve for most language pairs. On average, these are the best scores we have ever observed in one of our evaluation campaigns, including in WMT11, where results were inflated due to inclusion of the reference in the agreement rates. The results for intra-annotator agreement are more mixed: some improve greatly (Czech and German) while others degrade (French, Russian). Our special language, Finnish, also achieves very respectable scores. On average, again, we see the best intra-annotator agreement scores since WMT11.

It should be noted that the improvement is not caused by the "ties forced by our redundancy cleanup". If two systems A and F produced near-identical outputs, they are collapsed to one multi-system output AF and treated jointly in our agreement calculations, i.e. only in comparison with other outputs. It is only the final TrueSkill scores that include the tie $A = F$.

3.4 Producing the human ranking

The collected pairwise rankings are used to produce the official human ranking of the systems. For WMT14, we introduced a competition among multiple methods of producing this human ranking, selecting the method based on which could best predict the annotations in a portion of the collected pairwise judgments. The results of this competition were that (a) the competing metrics produced almost identical rankings across all tasks but that (b) one method, TrueSkill, had less variance across randomized runs, allowing us to make more confident cluster predictions. In light of these findings, this year, we produced the human ranking for each task using TrueSkill in the following fashion, following procedures adopted for WMT12: We produce 1,000 bootstrap-resampled runs over all of the available data. We then compute a *rank range* for each system by collecting the absolute rank of each system in each fold, throwing out the top and bottom 2.5%, and then clustering systems into equivalence classes containing systems with overlapping ranges, yielding a partial ordering over systems at the 95% confidence level.

The full list of the official human rankings for each task can be found in Table 6, which also reports all system scores, rank ranges, and clusters for all language pairs and all systems. The official interpretation of these results is that systems in the same cluster are considered tied. Given the large number of judgments that we collected, it was possible to group on average about two systems in a cluster, even though the systems in the middle are typically in larger clusters.

In Figure 4 and 5, we plotted the human evaluation result against everybody's favorite metric BLEU (some of the outlier online systems are

Language Pair	WMT11	WMT12	WMT13	WMT13$_r$	WMT13$_m$	WMT14	WMT15
Czech→English	0.400	0.311	0.244	0.342	0.279	0.305	0.458
English→Czech	0.460	0.359	0.168	0.408	0.075	0.360	0.438
German→English	0.324	0.385	0.299	0.443	0.324	0.368	0.423
English→German	0.378	0.356	0.267	0.457	0.239	0.427	0.423
French→English	0.402	0.272	0.275	0.405	0.321	0.357	0.343
English→French	0.406	0.296	0.231	0.434	0.237	0.302	0.317
Russian→English	—	—	0.278	0.315	0.324	0.324	0.372
English→Russian	—	—	0.243	0.416	0.207	0.418	0.336
Finnish→English	—	—	—	—	—	—	0.388
English→Finnish	—	—	—	—	—	—	0.549
Mean	0.395	0.330	0.260	0.403	0.251	0.367	0.405

Table 4: κ scores measuring inter-annotator agreement for WMT15. See Table 5 for corresponding intra-annotator agreement scores. WMT13$_r$ and WMT$_m$ refer to researchers' judgments and crowd-sourced judgments obtained using Mechanical Turk, respectively. WMT14 and WMT15 results are based on researchers' judgments only (hence, comparable to WMT13$_r$).

Language Pair	WMT11	WMT12	WMT13	WMT13$_r$	WMT13$_m$	WMT14	WMT15
Czech→English	0.597	0.454	0.479	0.483	0.478	0.382	0.694
English→Czech	0.601	0.390	0.290	0.547	0.242	0.448	0.584
German→English	0.576	0.392	0.535	0.643	0.515	0.344	0.801
English→German	0.528	0.433	0.498	0.649	0.452	0.576	0.676
French→English	0.673	0.360	0.578	0.585	0.565	0.629	0.510
English→French	0.524	0.414	0.495	0.630	0.486	0.507	0.426
Russian→English	—	—	0.450	0.363	0.477	0.629	0.506
English→Russian	—	—	0.513	0.582	0.500	0.570	0.492
Finnish→English	—	—	—	—	—	—	0.562
English→Finnish	—	—	—	—	—	—	0.697
Mean	0.583	0.407	0.479	0.560	0.464	0.522	0.595

Table 5: κ scores measuring intra-annotator agreement, i.e., self-consistency of judges, across for the past few years of the human evaluation campaign. Scores are much higher for WMT15 which makes sense as we enforce annotation consistency through our initial preprocessing which joins *near-identical translation candidates* into multi-system entries. It seems that the focus on actual differences in our annotation tasks as well as the possibility of having "easier" ranking scenarios for $n < 5$ candidate systems results in a higher annotator agreement, both for inter- and intra-annotator agreement scores.

not included to make the graphs viewable). The plots cleary suggest that a fair comparison of systems of different kinds cannot rely on automatic scores. Rule-based systems receive a much lower BLEU score than statistical systems (see for instance English–German, e.g., PROMT-RULE). The same is true to a lesser degree for statistical syntax-based systems (see English–German, UEDIN-SYNTAX) and online systems that were not tuned to the shared task (see Czech–English, CU-TECTO vs. the cluster of tuning task systems TT-*).

4 Quality Estimation Task

The fourth edition of the WMT shared task on quality estimation (QE) of machine translation (MT) builds on the previous editions of the task (Callison-Burch et al., 2012; Bojar et al., 2013, 2014), with tasks including both sentence and word-level estimation, using new training and test datasets, and an additional task: document-level prediction.

The goals of this year's shared task were:

- Advance work on sentence- and word-level quality estimation by providing larger datasets.
- Investigate the effectiveness of quality labels, features and learning methods for document-level prediction.
- Explore differences between sentence-level and document-level prediction.
- Analyse the effect of training data sizes and quality for sentence and word-level predic-

Czech–English

#	score	range	system
1	0.619	1	ONLINE-B
2	0.574	2	UEDIN-JHU
3	0.532	3-4	UEDIN-SYNTAX
	0.518	3-4	MONTREAL
4	0.436	5	ONLINE-A
5	-0.125	6	CU-TECTO
6	-0.182	7-9	TT-BLEU-MIRA-D
	-0.189	7-10	TT-ILLC-UVA
	-0.196	7-11	TT-BLEU-MERT
	-0.210	8-11	TT-AFRL
	-0.220	9-11	TT-USAAR-TUNA
7	-0.263	12-13	TT-DCU
	-0.297	13-15	TT-METEOR-CMU
	-0.320	13-15	TT-BLEU-MIRA-SP
	-0.320	13-15	TT-HKUST-MEANT
	-0.358	15-16	ILLINOIS

English–Czech

#	score	range	system
1	0.686	1	CU-CHIMERA
2	0.515	2-3	ONLINE-B
	0.503	2-3	UEDIN-JHU
3	0.467	4	MONTREAL
4	0.426	5	ONLINE-A
5	0.261	6	UEDIN-SYNTAX
6	0.209	7	CU-TECTO
7	0.114	8	COMMERCIAL1
8	-0.342	9-11	TT-DCU
	-0.342	9-11	TT-AFRL
	-0.346	9-11	TT-BLEU-MIRA-D
9	-0.373	12	TT-USAAR-TUNA
10	-0.406	13	TT-BLEU-MERT
11	-0.563	14	TT-METEOR-CMU
12	-0.808	15	TT-BLEU-MIRA-SP

Russian–English

#	score	range	system
1	0.494	1	ONLINE-G
2	0.311	2	ONLINE-B
3	0.129	3-6	PROMT-RULE
	0.116	3-6	AFRL-MIT-PB
	0.113	3-6	AFRL-MIT-FAC
	0.104	3-7	ONLINE-A
	0.051	6-8	AFRL-MIT-H
	0.010	7-10	LIMSI-NCODE
	-0.021	8-10	UEDIN-SYNTAX
	-0.031	8-10	UEDIN-JHU
4	-0.218	11	USAAR-GACHA
5	-0.278	12	USAAR-GACHA
6	-0.781	13	ONLINE-F

German–English

#	score	range	system
1	0.567	1	ONLINE-B
2	0.319	2-3	UEDIN-JHU
	0.298	2-4	ONLINE-A
	0.258	3-5	UEDIN-SYNTAX
	0.228	4-5	KIT
3	0.141	6-7	RWTH
	0.095	6-7	MONTREAL
4	-0.172	8-10	ILLINOIS
	-0.177	8-10	DFKI
	-0.221	9-10	ONLINE-C
5	-0.304	11	ONLINE-F
6	-0.489	12-13	MACAU
	-0.544	12-13	ONLINE-E

French–English

#	score	range	system
1	0.498	1-2	ONLINE-B
	0.446	1-3	LIMSI-CNRS
	0.415	1-3	UEDIN-JHU
2	0.275	4-5	MACAU
	0.223	4-5	ONLINE-A
3	-0.423	6	ONLINE-F
4	-1.434	7	ONLINE-E

English–French

#	score	range	system
1	0.540	1	LIMSI-CNRS
2	0.304	2-3	ONLINE-A
	0.258	2-4	UEDIN-JHU
	0.215	3-4	ONLINE-B
3	-0.001	5	CIMS
4	-0.338	6	ONLINE-F
5	-0.977	7	ONLINE-E

English–Russian

#	score	range	system
1	1.015	1	PROMT-RULE
2	0.521	2	ONLINE-G
3	0.217	3	ONLINE-B
4	0.122	4-5	LIMSI-NCODE
	0.075	4-5	ONLINE-A
5	0.014	6	UEDIN-JHU
6	-0.138	7	UEDIN-SYNTAX
7	-0.276	8	USAAR-GACHA
8	-0.333	9	USAAR-GACHA
9	-1.218	10	ONLINE-F

English–German

#	score	range	system
1	0.359	1-2	UEDIN-SYNTAX
	0.334	1-2	MONTREAL
2	0.260	3-4	PROMT-RULE
	0.235	3-4	ONLINE-A
3	0.148	5	ONLINE-B
4	0.086	6	KIT-LIMSI
5	0.036	7-9	UEDIN-JHU
	0.003	7-11	ONLINE-F
	-0.001	7-11	ONLINE-C
	-0.018	8-11	KIT
	-0.035	9-11	CIMS
6	-0.133	12-13	DFKI
	-0.137	12-13	ONLINE-E
7	-0.235	14	UDS-SANT
8	-0.400	15	ILLINOIS
9	-0.501	16	IMS

Finnish–English

#	score	range	system
1	0.675	1	ONLINE-B
2	0.280	2-4	PROMT-SMT
	0.246	2-5	ONLINE-A
	0.236	2-5	UU
	0.182	4-7	UEDIN-JHU
	0.160	5-7	ABUMATRAN-COMB
	0.144	5-8	UEDIN-SYNTAX
	0.081	7-8	ILLINOIS
3	-0.081	9	ABUMATRAN-HFS
4	-0.177	10	MONTREAL
5	-0.275	11	ABUMATRAN
6	-0.438	12-13	LIMSI
	-0.513	13-14	SHEFFIELD
	-0.520	13-14	SHEFF-STEM

English–Finnish

#	score	range	system
1	1.069	1	ONLINE-B
2	0.548	2	ONLINE-A
3	0.210	3	UU
4	0.042	4	ABUMATRAN-COMB
5	-0.059	5	ABUMATRAN-COMB
6	-0.143	6-7	AALTO
	-0.184	6-8	UEDIN-SYNTAX
	-0.212	6-8	ABUMATRAN
7	-0.342	9	CMU
8	-0.929	10	CHALMERS

Table 6: Official results for the WMT15 translation task. Systems are ordered by their inferred system means, though systems within a cluster are considered tied. Lines between systems indicate clusters according to bootstrap resampling at p-level $p \leq .05$. Systems with grey background indicate use of resources that fall outside the constraints provided for the shared task.

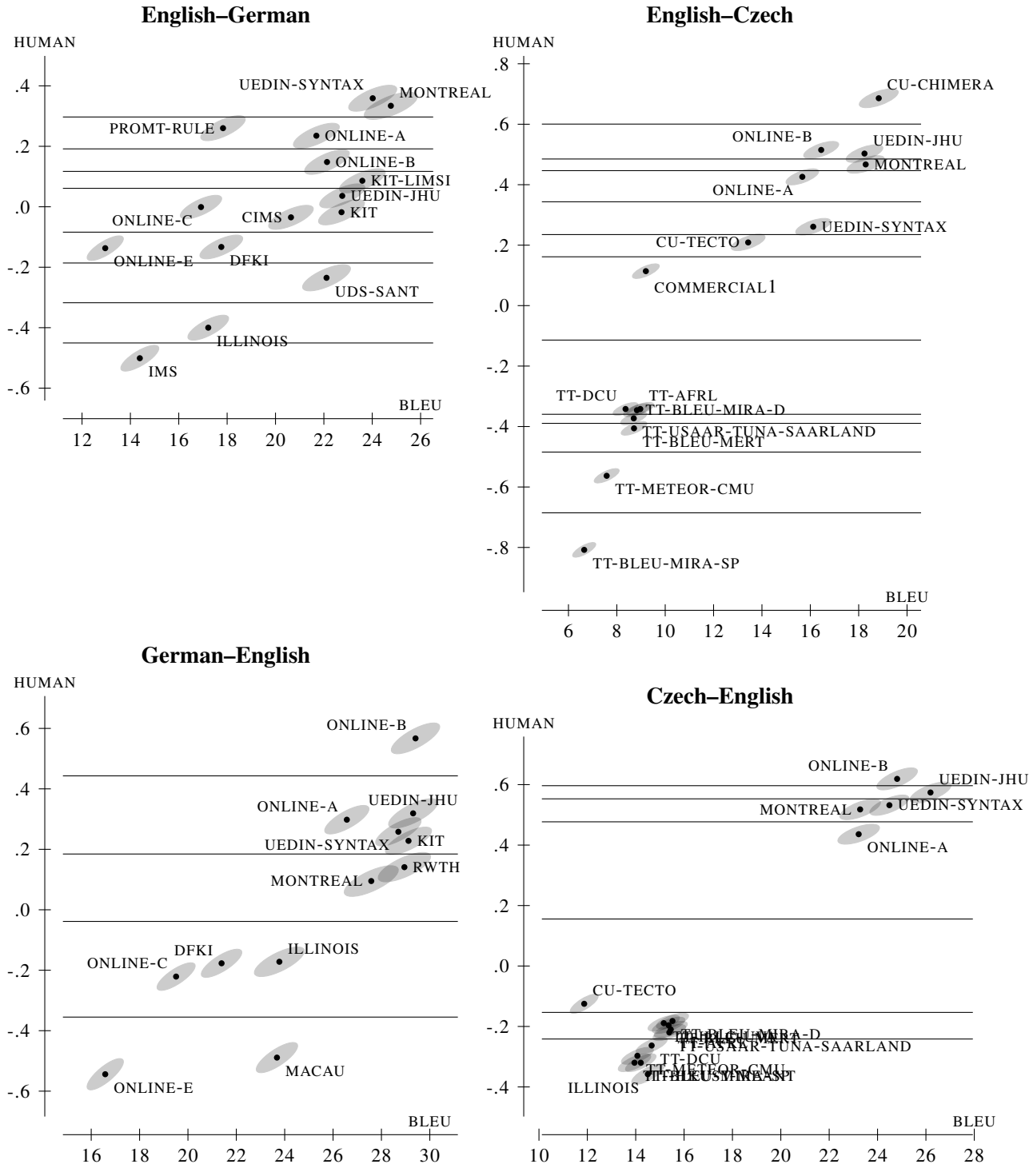

Figure 4: Human evaluation scores versus BLEU scores for the German–English and Czech–English language pairs illustrate the need for human evaluation when comparing systems of different kind. Confidence intervals are indicated by the shaded ellipses. Rule-based systems and to a lesser degree syntax-based statistical systems receive a lower BLEU score than their human score would indicate. The big cluster in the Czech-English plot are tuning task submissions.

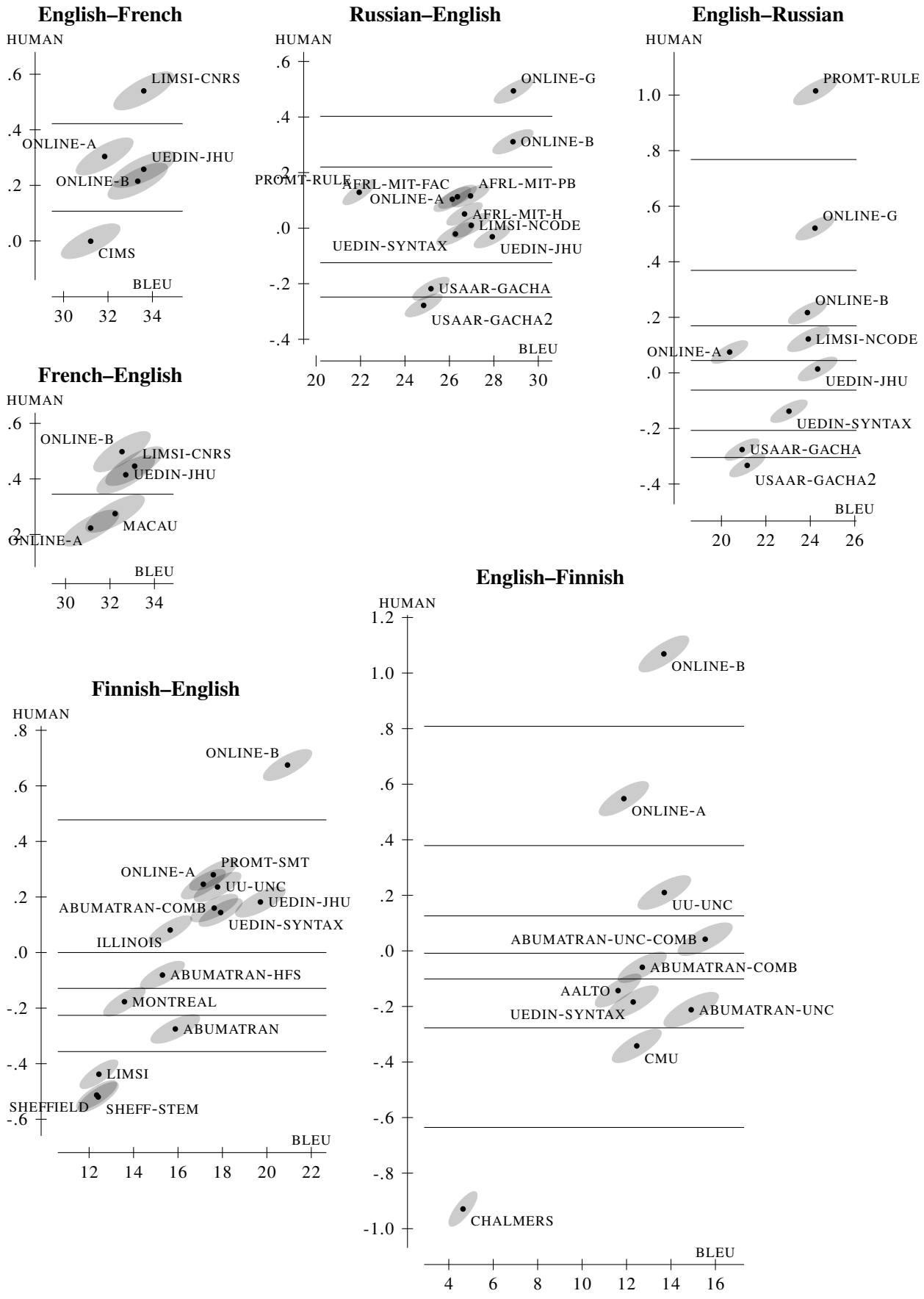

Figure 5: Human evaluation versus BLEU scores for the French–English, Russian–English, and Finnish-English language pairs.

tion, particularly the use of annotations obtained from crowdsourced post-editing.

Three tasks were proposed: Task 1 at sentence level (Section 4.3), Task 2 at word level (Section 4.4), and Task 3 at document level (Section 4.5). Tasks 1 and 2 provide the same dataset with English-Spanish translations generated by the statistical machine translation (SMT) system, while Task 3 provides two different datasets, for two language pairs: English-German (EN-DE) and German-English (DE-EN) translations taken from all participating systems in WMT13 (Bojar et al., 2013). These datasets were annotated with different labels for quality: for Tasks 1 and 2, the labels were automatically derived from the post-editing of the machine translation output, while for Task 3, scores were computed based on reference translations using Meteor (Banerjee and Lavie, 2005). Any external resource, including additional quality estimation training data, could be used by participants (no distinction between *open* and *close* tracks was made). As presented in Section 4.1, participants were also provided with a baseline set of features for each task, and a software package to extract these and other quality estimation features and perform model learning, with suggested methods for all levels of prediction. Participants, described in Section 4.2, could submit up to two systems for each task.

Data used to build MT systems or internal system information (such as model scores or n-best lists) were not made available this year as multiple MT systems were used to produce the datasets, especially for Task 3, including online and rule-based systems. Therefore, as a general rule, participants could only use black-box features.

4.1 Baseline systems

Sentence-level baseline system: For Task 1, QUEST[7] (Specia et al., 2013) was used to extract 17 MT system-independent features from the source and translation (target) files and parallel corpora:

- Number of tokens in the source and target sentences.
- Average source token length.
- Average number of occurrences of the target word within the target sentence.
- Number of punctuation marks in source and target sentences.
- Language model (LM) probability of source and target sentences based on models for the WMT News Commentary corpus.
- Average number of translations per source word in the sentence as given by IBM Model 1 extracted from the WMT News Commentary parallel corpus, and thresholded such that $P(t|s) > 0.2/P(t|s) > 0.01$.
- Percentage of unigrams, bigrams and trigrams in frequency quartiles 1 (lower frequency words) and 4 (higher frequency words) in the source language extracted from the WMT News Commentary corpus.
- Percentage of unigrams in the source sentence seen in the source side of the WMT News Commentary corpus.

These features were used to train a Support Vector Regression (SVR) algorithm using a Radial Basis Function (RBF) kernel within the SCIKIT-LEARN toolkit.[8] The γ, ϵ and *C* parameters were optimised via grid search with 5-fold cross validation on the training set. We note that although the system is referred to as "baseline", it is in fact a strong system. It has proved robust across a range of language pairs, MT systems, and text domains for predicting various forms of post-editing effort (Callison-Burch et al., 2012; Bojar et al., 2013, 2014).

Word-level baseline system: For Task 2, the baseline features were extracted with the MARMOT tool[9]. For the baseline system we used a number of features that have been found the most informative in previous research on word-level quality estimation. Our baseline set of features is loosely based on the one described in (Luong et al., 2014). It contains the following 25 features:

- Word count in the source and target sentences, source and target token count ratio. Although these features are sentence-level (i.e. their values will be the same for all words in a sentence), but the length of a sentence might influence the probability of a word being incorrect.

[7] https://github.com/lspecia/quest

[8] http://scikit-learn.org/

[9] https://github.com/qe-team/marmot

- Target token, its left and right contexts of one word.
- Source token aligned to the target token, its left and right contexts of one word. The alignments were produced with the `force_align.py` script, which is part of **cdec** (Dyer et al., 2010). It allows to align new parallel data with a pre-trained alignment model built with the cdec word aligner (**fast_align**). The alignment model was trained on the Europarl corpus (Koehn, 2005).
- Boolean dictionary features: whether target token is a stopword, a punctuation mark, a proper noun, a number.
- Target language model features:
  - The order of the highest order n-gram which starts or ends with the target token.
  - Backoff behaviour of the n-grams (t_{i-2}, t_{i-1}, t_i), (t_{i-1}, t_i, t_{i+1}), (t_i, t_{i+1}, t_{i+2}), where t_i is the target token (the backoff behaviour is computed as described in (Raybaud et al., 2011)).
- The order of the highest order n-gram which starts or ends with the source token.
- Boolean pseudo-reference feature: 1 if the token is contained in a pseudo-reference, 0 otherwise. The pseudo-reference used for this feature is the automatic translation generated by an English-Spanish phrase-based SMT system trained on the Europarl corpus with standard settings.[10]
- The part-of-speech tags of the target and source tokens.
- The number of senses of the target and source tokens in WordNet.

We model the task as a sequence prediction problem and train our baseline system using the Linear-Chain Conditional Random Fields (CRF) algorithm with the CRF++ tool.[11]

Document-level baseline system: For Task 3, the baseline features for sentence-level prediction were used. These are aggregated by summing or averaging their values for the entire document. Features that were summed: number of tokens in the source and target sentences and number of punctuation marks in source and target sentences. All other features were averaged. The implementation for document-level feature extraction is available in QUEST++ (Specia et al., 2015).[12]

These features were then used to train a SVR algorithm with RBF kernel using the SCIKIT-LEARN toolkit. The γ, ϵ and *C* parameters were optimised via grid search with 5-fold cross validation on the training set.

4.2 Participants

Table 7 lists all participating teams submitting systems to any of the tasks. Each team was allowed up to two submissions for each task and language pair. In the descriptions below, participation in specific tasks is denoted by a task identifier.

DCU-SHEFF (Task 2): The system uses the baseline set of features provided for the task. Two pre-processing data manipulation techniques were used: data selection and data bootstrapping. Data selection filters out sentences which have the smallest proportion of erroneous tokens and are assumed to be the least useful for the task. Data bootstrapping enhances the training data with incomplete training sentences (e.g. the first k words of a sentence of the length N, where $k < N$). This technique creates additional data instances and boosts the importance of errors occurring in the training data. The combination of these techniques doubled the F_1 score for the "BAD" class, as compared to a models trained on the entire dataset given for the task. The labelling was performed with a CRF model trained using the CRF++ tool, as in the baseline system.

HDCL (Task 2): HDCL's submissions are based on a deep neural network that learns continuous feature representations from scratch, i.e. from bilingual contexts. The network was pre-trained by initialising the word lookup-table with distributed word representations,

[10] `http://www.statmt.org/moses/?n=Moses.Baseline`

[11] `http://taku910.github.io/crfpp/`

[12] `https://github.com/ghpaetzold/questplusplus`

ID	Participating team
DCU-SHEFF	Dublin City University, Ireland and University of Sheffield, UK (Logacheva et al., 2015)
HDCL	Heidelberg University, Germany (Kreutzer et al., 2015)
LORIA	Lorraine Laboratory of Research in Computer Science and its Applications, France (Langlois, 2015)
RTM-DCU	Dublin City University, Ireland (Bicici et al., 2015)
SAU-KERC	Shenyang Aerospace University, China (Shang et al., 2015)
SHEFF-NN	University of Sheffield Team 1, UK (Shah et al., 2015)
UAlacant	Alicant University, Spain (Esplà-Gomis et al., 2015a)
UGENT	Ghent University, Belgium (Tezcan et al., 2015)
USAAR-USHEF	University of Sheffield, UK and Saarland University, Germany (Scarton et al., 2015a)
USHEF	University of Sheffield, UK (Scarton et al., 2015a)
HIDDEN	Undisclosed

Table 7: Participants in the WMT15 quality estimation shared task.

and fine-tuned for the quality estimation classification task by back-propagating word-level prediction errors using stochastic gradient descent. In addition to the continuous space deep model, a shallow linear classifier was trained on the provided baseline features and their quadratic expansion. One of the submitted systems (QUETCH) relies on the deep model only, the other (QUETCHPLUS) is a linear combination of the QUETCH system score, the linear classifier score, and binary and binned baseline features. The system combination yielded significant improvements, showing that the deep and shallow models each contributes complementary information to the combination.

LORIA (Task 1): The LORIA system for Task 1 is based on a standard machine learning approach where source-target sentences are described by numerical vectors and SVR is used to learn a regression model between these vectors and quality scores. Feature vectors used the 17 baseline features, two Latent Semantic Indexing (LSI) features and 31 features based on pseudo-references. The LSI approach considers source-target pairs as documents, and projects the TF-IDF words-documents matrix into a reduced numerical space. This leads to a measure of similarity between a source and a target sentence, which was used as a feature. Two of these features were used based on two matrices, one from the Europarl corpus and one from the official training data. Pseudo-references were produced by three online systems. These features measure the intersection between n-gram sets of the target sentence and of the pseudo-references. Three sets of features were extracted from each online system, and a fourth feature was extracted measuring the inter-agreement among the three online systems and the target system.

RTM-DCU (Tasks 1, 2, 3): RTM-DCU systems are based on referential translation machines (RTM) (Biçici, 2013; Biçici and Way, 2014). RTMs propose a language independent approach and avoid the need to access any task- or domain-specific information or resource. The submissions used features that indicate the closeness between instances to the available training data, the difficulty of translating them, and the presence of acts of translation for data transformation. SVR was used for document and sentence-level prediction tasks, also in combination with feature selection or partial least squares, and global linear models with dynamic learning were used for the word-level prediction task.

SAU (Task 2): The SAU submissions used a CRF model to predict the binary labels for Task 2. They rely on 12 basic features and 85 combination features. The ratio between OK and BAD labels was found to be 4:1 in the training set. Two strategies were proposed to

solve this problem of label ratio imbalance. The first strategy is to replace "OK" labels with sub-labels to balance label distribution, where the sub-labels are OK_B, OK_I, OK_E, OK (depending on the position of the token in the sentence). The second strategy is to reconstruct the training set to include more "BAD" words.

SHEFF-NN (Tasks 1, 2): SHEFF-NN submissions were based on (i) a Continuous Space Language Model (CSLM) to extract additional features for Task 1 (SHEF-GP and SHEF-SVM), (ii) a Continuous Bag-of-Words (CBOW) model to produce word embeddings as features for Task 2 (SHEF-W2V), and (iii) a combination of features produced by QUEST++ and a feature produced with word embedding models (SHEF-QuEst++). SVR and Gaussian Processes were used to learn prediction models for Task 1, and a CRF algorithm for binary tagging models in Task 2 (Pystruct Linear-chain CRF trained with a structured SVM for system SHEF-W2V, and CRFSuite Adaptive Regularisation of Weight Vector (AROW) and Passive Aggressive (PA) algorithms for system SHEF-QuEst++). Interesting findings for Task 1 were that (i) CSLM features always bring improvements whenever added to either baseline or complete feature sets and (ii) CSLM features alone perform better than the baseline features. For Task 2, the results obtained by SHEF-W2V are promising: although it uses only features learned in unsupervised fashion (CBOW word embeddings), it was able to outperform the baseline as well as many other systems. Further, combining the source-to-target cosine similarity feature with the ones produced by QUEST++ led to improvements in the F_1 of "BAD" labels.

UAlacant (Task 2): The submissions of the Universitat d'Alacant team were obtained by applying the approach in (Esplà-Gomis et al., 2015b), which uses any source of bilingual information available as a black-box in order to spot sub-segment correspondences between a sentence S in the source language and a given translation hypothesis T in the target language. These sub-segment correspondences are used to extract a collection of features that is then used by a multilayer perceptron to determine the word-level predicted score. Three sources of bilingual information available online were used: two online machine translation systems, Apertium[13] and Google Translate; and the bilingual concordancer Reverso Context.[14] Two submissions were made for Task 2: one using only the 70 features described in (Esplà-Gomis et al., 2015b), and one combining them with the baseline features provided by the task organisers.

UGENT (Tasks 1, 2): The submissions for the word-level task used 55 new features in combination with the baseline feature set to train binary classifiers. The new features try to capture either accuracy (meaning transfer from source to target sentence) using word and phrase alignments, or fluency (well-formedness of target sentence) using language models trained on word surface forms and on part-of-speech tags. Based on the combined feature set, SCATE-MBL uses a memory-based learning (MBL) algorithm for binary classification. SCATE-HYBRID uses the same feature set and forms a classifier ensemble using CRFs in combination with the MBL system for predicting word-level quality. For the sentence-level task, SCATE-SVM-single uses a single feature to train SVR models, which is based on the percentage of words that are labelled as "BAD" by the word-level quality estimation system SCATE-HYBRID. SCATE-SVM adds 16 new features to this single feature and the baseline feature set to train SVR models using an RBF kernel. Additional language resources are used to extract the new features for both tasks.

USAAR-USHEF (Task 3): The systems submitted for both EN-DE and DE-EN (called BFF) were built by using a exhaustive search for feature selection over the official baseline features. In order to select the best features, a Bayesian Ridge classifier was trained for each feature combination and the classifiers were evaluated in terms of Mean Average Error (MAE): the classifier with the smallest

[13] http://www.apertium.org

[14] http://context.reverso.net/translation/

MAE was considered the best. For EN-DE, the selected features were: average source token length, percentage of unigrams and of trigrams in fourth quartile of frequency in a corpus of the source language. For DE-EN, the best features were: number of occurrences of the target word within the target hypothesis, percentage of unigrams and of trigrams in first quartile of frequency in a corpus of the source language. This provide an indication of which features of the baseline set contribute for document-level quality estimation.

USHEF (Task 3): The system submitted for the EN-DE document-level task was built by using the 17 official baseline features, plus discourse features (repetition of words, lemmas and nouns and ratio of repetitions – as implemented in QUEST++. For DE-EN, a combination of the 17 baseline features, the discourse repetition features and discourse-aware features extracted from syntactic and discourse parsers was used. The new discourse features are: number of pronouns, number of connectives, number of satellite and nucleus relations in the RST (Rhetorical Structure Theory) tree for the document and number of EDU (Elementary Discourse Units) breaks in the text. A backward feature selection approach, based on the feature rank of SCIKIT-LEARN's Random Forest implementation, was also applied. For both languages pairs, the same algorithm as that of the baseline system was used: the SCIKIT-LEARN implementation of SVR with RBF kernel and hyper-parameters optimised via grid-search.

HIDDEN (Task 3): This submission, whose creators preferred to remain anonymous, estimates the quality of a given document by explicitly identifying potential translation errors in it. Translation error detection is implemented as a combination of human expert knowledge and different language processing tools, including named entity recognition, part-of-speech tagging and word alignments. In particular, the system looks for patterns of errors defined by human experts, taking into account the actual words and the additional linguistic information. With this approach, a wide variety of errors can be detected: from simple misspellings and typos to complex lack of agreement (in genre, number and tense), or lexical inconsistencies. Each error category is assigned an "importance", again according to human knowledge, and the amount of error in the document is computed as the weighted sum of the identified errors. Finally, the documents are sorted according to this figure to generate the final submission to the ranking variant of Task 3.

4.3 Task 1: Predicting sentence-level quality

This task consists in scoring (and ranking) translation sentences according to the percentage of their words that need to be fixed. It is similar to Task 1.2 in WMT14. HTER (Snover et al., 2006b) is used as quality score, i.e. the minimum edit distance between the machine translation and its manually post-edited version in [0,1].

As in previous years, two variants of the results could be submitted:

- **Scoring**: An absolute HTER score for each sentence translation, to be interpreted as an error metric: lower scores mean better translations.
- **Ranking**: A ranking of sentence translations for all source sentences from best to worst. For this variant, it does not matter how the ranking is produced (from HTER predictions or by other means). The reference ranking is defined based on the true HTER scores.

Data The data is the same as that used for the WMT15 Automatic Post-editing task,[15] as kindly provided by Unbabel.[16] Source segments are tokenized English sentences from the news domain with at least four tokens. Target segments are tokenized Spanish translations produced by an online SMT system. The human post-editions are a manual revision of the target, collected using Unbabel's crowd post-editing platform. HTER labels were computed using the TERCOM tool[17] with default settings (tokenised, case insensitive, exact matching only), but with scores capped to 1.

As training and development data, we provided English-Spanish datasets with $11,271$ and $1,000$ source sentences, their machine translations, post-editions and HTER scores, respectively. As test data, we provided an additional

[15] http://www.statmt.org/wmt15/ape-task.html
[16] https://unbabel.com/
[17] http://www.cs.umd.edu/~snover/tercom/

set of $1,817$ English-Spanish source-translations pairs produced by the same MT system used for the training data.

Evaluation Evaluation was performed against the true HTER label and/or ranking, using the same metrics as in previous years:

- Scoring: Mean Average Error (MAE) (primary metric, official score for ranking submissions), Root Mean Squared Error (RMSE).
- Ranking: DeltaAvg (primary metric) and Spearman's ρ rank correlation.

Additionally, we included Pearson's r correlation against the true HTER label, as suggested by Graham (2015).

Statistical significance on MAE and DeltaAvg was computed using a pairwise bootstrap resampling (1K times) approach with 95% confidence intervals. [18] For Pearson's r correlation, we measured significance using the Williams test, as also suggested in (Graham, 2015).

Results Table 8 summarises the results for the ranking variant of Task 1. They are sorted from best to worst using the DeltaAvg metric scores as primary key and the Spearman's ρ rank correlation scores as secondary key.

The results for the scoring variant are presented in Table 9, sorted from best to worst by using the MAE metric scores as primary key and the RMSE metric scores as secondary key.

Pearson's r coefficients for all systems against HTER is given in Table 10. As discussed in (Graham, 2015), the results according to this metric can rank participating systems differently. In particular, we note the SHEF/GP submission, are which is deemed significantly worse than the baseline system according to MAE, but substantially better than the baseline according to Pearson's correlation. Graham (2015) argues that the use of MAE as evaluation score for quality estimation tasks is inadequate, as MAE is very sensitive to variance. This means that a system that outputs predictions with high variance is more likely to have high MAE score, even if the distribution follows that of the true labels. Interestingly, according to Pearson's correlation, the systems are ranked exactly in the same way as according to our DeltaAvg metric. The only difference is that the 4th place is now considered significantly different from the three winning submissions. She also argues that the significance tests used with MAE, based on randomised resampling, assume that the data is independent, which is not the case. Therefore, we apply the suggested Williams significance test for this metric.

[18] http://www.quest.dcs.shef.ac.uk/wmt15_files/bootstrap-significance.pl

4.4 Task 2: Predicting word-level quality

The goal of this task is to evaluate the extent to which we can detect word-level errors in MT output. Often, the overall quality of a translated segment is significantly harmed by specific errors in a small proportion of words. Various classes of errors can be found in translations, but for this task we consider all error types together, aiming at making a binary distinction between 'GOOD' and 'BAD' tokens. The decision to bucket all error types together was made because of the lack of sufficient training data that could allow consideration of more fine-grained error tags.

Data This year's word-level task uses the same dataset as Task 1, for a single language pair: English-Spanish. Each instance of the training, development and test sets consists of the following elements:

- Source sentence (English).
- Automatic translation (Spanish).
- Manual post-edition of the automatic translation.
- Word-level binary ("OK"/"BAD") labelling of the automatic translation.

The binary labels for the datasets were acquired automatically with the TERCOM tool (Snover et al., 2006b).[19] This tool computes the edit distance between machine-translated sentence and its reference (in this case, its post-edited version). It identifies four types of errors: *substitution* of a word with another word, *deletion* of a word (word was omitted by the translation system), *insertion* of a word (a redundant word was added by the translation system), and word or sequence of words *shift* (word order error). Every word in the machine-translated sentence is tagged with one of these error types or not tagged if it matches a word from the reference.

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

System ID	**DeltaAvg** ↑	Spearman's ρ ↑
English-Spanish		
• LORIA/17+LSI+MT+FILTRE	6.51	0.36
• LORIA/17+LSI+MT	6.34	0.37
• RTM-DCU/RTM-FS+PLS-SVR	6.34	0.37
• RTM-DCU/RTM-FS-SVR	6.09	0.35
UGENT-LT3/SCATE-SVM	6.02	0.34
UGENT-LT3/SCATE-SVM-single	5.12	0.30
SHEF/SVM	5.05	0.28
SHEF/GP	3.07	0.28
Baseline SVM	2.16	0.13

Table 8: Official results for the ranking variant of the WMT15 quality estimation Task 1. The winning submissions are indicated by a •. These are the top-scoring submission and those that are not significantly worse according to pairwise bootstrap resampling (1K times) with 95% confidence intervals. The systems in the gray area are not different from the baseline system at a statistically significant level according to the same test.

System ID	**MAE** ↓	RMSE ↓
English-Spanish		
• RTM-DCU/RTM-FS+PLS-SVR	13.25	17.48
• LORIA/17+LSI+MT+FILTRE	13.34	17.35
• RTM-DCU/RTM-FS-SVR	13.35	17.68
• LORIA/17+LSI+MT	13.42	17.45
• UGENT-LT3/SCATE-SVM	13.71	17.45
UGENT-LT3/SCATE-SVM-single	13.76	17.79
SHEF/SVM	13.83	18.01
Baseline SVM	14.82	19.13
SHEF/GP	15.16	18.97

Table 9: Official results for the scoring variant of the WMT15 quality estimation Task 1. The winning submissions are indicated by a •. These are the top-scoring submission and those that are not significantly worse according to bootstrap resampling (1K times) with 95% confidence intervals. The systems in the gray area are not different from the baseline system at a statistically significant level according to the same test.

System ID	**Pearson's** r ↑
• LORIA/17+LSI+MT+FILTRE	0.39
• LORIA/17+LSI+MT	0.39
• RTM-DCU/RTM-FS+PLS-SVR	0.38
RTM-DCU/RTM-FS-SVR	0.38
UGENT-LT3/SCATE-SVM	0.37
UGENT-LT3/SCATE-SVM-single	0.32
SHEF/SVM	0.29
SHEF/GP	0.19
Baseline SVM	0.14

Table 10: Alternative results for the scoring variant of the WMT15 quality estimation Task 1. The winning submissions are indicated by a •. These are the top-scoring submission and those that are not significantly worse according to Williams test with 95% confidence intervals. The systems in the gray area are not different from the baseline system at a statistically significant level according to the same test.

All the untagged (correct) words were tagged with "OK", while the words tagged with substitution and insertion errors were assigned the tag "BAD". The deletion errors are not associated with any word in the automatic translation, so we could not consider them. We also disabled the shift errors by running TERCOM with the option '-d 0'. The reason for that is the fact that searching for shifts introduces significant noise in the annotation. The system cannot discriminate be-

tween cases where a word was really shifted and where a word (especially common words such as prepositions, articles and pronouns) was deleted in one part of the sentence and then independently inserted in another part of this sentence, i.e. to correct an unrelated error. The statistics of the datasets are outlined in Table 11.

	Sentences	Words	% of "BAD" words
Training	11,271	257,548	19.14
Dev	1,000	23,207	19.18
Test	1,817	40,899	18.87

Table 11: Datasets for Task 2.

Evaluation Submissions were evaluated in terms of classification performance against the original labels. The main evaluation metric is the average F_1 for the "BAD" class. Statistical significance on F_1 for the "BAD" class was computed using approximate randomization tests.[20]

Results The results for Task 2 are summarised in Table 12. The results are ordered by F_1 score for the error (BAD) class.

Using the F_1 score for the word-level estimation task has a number of drawbacks. First of all, we cannot use it as the single metric to evaluate the system's quality. The F_1 score of the class "BAD" becomes an inadequate metric when one is also interested in the tagging of correct words. In fact, a naive baseline which tags all words with the class "BAD" would yield 31.75 F_1 score for the "BAD" class in the test set of this task, which is close to some of the submissions and by far exceeds the baseline, although this tagging is uninformative.

We could instead use the weighted F_1 score, which would lead to a single F_1 figure where every class is given a weight according to its frequency in the test set. However, we believe the weighted F_1 score does not reflect the real quality of the systems either. Since there are many more instances of the "GOOD" class than there are of the "BAD" class, the performance on the "BAD" class does not contribute much weight to the overall score, and changes in accuracy of error prediction on this less frequent class can go unnoticed. The weighted F_1 score for the strategy which tags all words as "GOOD" would be 72.66, which is higher than the score of many submissions. However, similar to the case of tagging all words as "BAD", this strategy is uninformative. In an attempt to find more intuitive ways of evaluating word-level tasks, we introduce a new metric called *sequence correlation*. It gives higher importance to the instances of the "BAD" class and is robust against uninformative tagging.

The basis of the sequence correlation metric is the number of matching labels in the reference and the hypothesis, analogously to a precision metric. However, it has some additional features that are aimed at making it more reliable. We consider the tagging of each sentence separately as a sequence of tags. We divide each sequence into sub-sequences tagged by the same tag, for example, the sequence "OK BAD OK OK OK" will be represented as a list of 3 sub-sequences: ["OK", "BAD", "OK OK OK"]. Each subsequence has also the information on its position in the original sentence. The sub-sequences of the reference and the hypothesis are then intersected, and the number of matching tags in the corresponding subsequences is computed so that every sub-sequence can be used only once. Let us consider the following example:

```
Reference:   OK  BAD  OK  OK  OK
Hypothesis:  OK   OK  OK  OK  OK
```

Here, the reference has three sub-sequences, as in the previous example, and the hypothesis consists of only one sub-sequence which coincides with the hypothesis itself, because all the words were tagged with the "OK" label. The precision score for this sentence will be 0.8, as 4 of 5 labels match in this example. However, we notice that the hypothesis sub-sequence covers two matching sub-sequences of the reference: word 1 and words 3–5. According to our metric, the hypothesis sub-sequence can be used for the intersection only once, giving either 1 of 5 or 3 of 5 matching words. We choose the highest value and get the score of 0.6. Thus, the intersection procedure downweighs the uninformative hypotheses where all words are tagged with one tag.

In order to compute the sequence correlation we need to get the set of spans for each label in both the prediction and the reference, and then intersect them. A set of spans of each tag t in the string $\mathbf{w}$ is computed as follows:

[20] http://www.nlpado.de/~sebastian/software/sigf.shtml

System ID	weighted F_1 All	F_1 **Bad** ↑	F_1 GOOD
English-Spanish			
• UAlacant/OnLine-SBI-Baseline	71.47	43.12	78.07
• HDCL/QUETCHPLUS	72.56	43.05	79.42
UAlacant/OnLine-SBI	69.54	41.51	76.06
SAU/KERC-CRF	77.44	39.11	86.36
SAU/KERC-SLG-CRF	77.4	38.91	86.35
SHEF2/W2V-BI-2000	65.37	38.43	71.63
SHEF2/W2V-BI-2000-SIM	65.27	38.40	71.52
SHEF1/QuEst++-AROW	62.07	38.36	67.58
UGENT/SCATE-HYBRID	74.28	36.72	83.02
DCU-SHEFF/BASE-NGRAM-2000	67.33	36.60	74.49
HDCL/QUETCH	75.26	35.27	84.56
DCU-SHEFF/BASE-NGRAM-5000	75.09	34.53	84.53
SHEF1/QuEst++-PA	26.25	34.30	24.38
UGENT/SCATE-MBL	74.17	30.56	84.32
RTM-DCU/s5-RTM-GLMd	76.00	23.91	88.12
RTM-DCU/s4-RTM-GLMd	75.88	22.69	88.26
Baseline	75.31	16.78	88.93

Table 12: Official results for the WMT15 quality estimation Task 2. The winning submissions are indicated by a •. These are the top-scoring submission and those that are not significantly worse according to approximate randomization tests with 95% confidence intervals. Submissions whose results are statistically different from others according to the same test are grouped by a horizontal line.

$$S_t(\mathbf{w}) = \{\mathbf{w}_{[b:e]}\},\ \forall i\ s.t.\ b \leqslant i \leqslant e : w_i = t$$

where $\mathbf{w}_{[b:e]}$ is a substring $w_b, w_{b+1}, ..., w_{e-1}, w_e$. Then the intersection of spans for all labels is:

$$Int(\mathbf{y}, \hat{\mathbf{y}}) = \sum_{t \in \{0;1\}} \lambda_t \sum_{s_{\mathbf{y}} \in S_t(\mathbf{y})} \sum_{s_{\hat{\mathbf{y}}} \in S_t(\hat{\mathbf{y}})} |s_{\mathbf{y}} \cap s_{\hat{\mathbf{y}}}|$$

Here λ_t is the weight of a tag t in the overall result. It is inversely proportional the number of instances of this tag in the reference:

$$\lambda_t = \frac{|\mathbf{y}|}{c_t(\hat{\mathbf{y}})}$$

where $c_t(\hat{\mathbf{y}})$ is the number of words labelled with the label t in the prediction. Thus we give the equal importance to all tags.

The sum of matching spans is also weighted by the ratio of the number of spans in the hypothesis and the reference. This is done to downweigh the system tagging if the number of its spans differs from the number of spans provided in the gold standard. This ratio is computed as follows:

$$r(\mathbf{y}, \hat{\mathbf{y}}) = \min(\frac{|\mathbf{y}|}{|\hat{\mathbf{y}}|}; \frac{|\hat{\mathbf{y}}|}{|\mathbf{y}|})$$

This ratio is 1 if the number of spans is equal for the hypothesis and the reference, and less than 1 otherwise.

The final score for a sentence is produced as follows:

$$SeqCor(\mathbf{y}, \hat{\mathbf{y}}) = \frac{r(\mathbf{y}, \hat{\mathbf{y}}) \cdot Int(\mathbf{y}, \hat{\mathbf{y}})}{|\mathbf{y}|} \quad (1)$$

Then the overall sequence correlation for the whole dataset is the average of sentence scores.

Table 13 shows the results of the evaluation according to the sequence correlation metric. The results for the two metrics are quite different: one of the highest scoring submissions according to the F_1-BAD score is only the third under the sequence correlation metric, and vice versa: the submissions with the highest sequence correlation feature in 3rd place according to F_1-BAD score. However, the system rankings produced by two metrics are correlated — their Spearman's correlation coefficient between them is 0.65.

System ID	Sequence Correlation
English-Spanish	
• SAU/KERC-CRF	34.22
• SAU/KERC-SLG-CRF	34.09
• UAlacant/OnLine-SBI-Baseline	33.84
UAlacant/OnLine-SBI	32.81
HDCL/QUETCH	32.13
HDCL/QUETCHPLUS	31.38
DCU-SHEFF/BASE-NGRAM-5000	31.23
UGENT/SCATE-HYBRID	30.15
DCU-SHEFF/BASE-NGRAM-2000	29.94
UGENT/SCATE-MBL	28.43
SHEF2/W2V-BI-2000	27.65
SHEF2/W2V-BI-2000-SIM	27.61
SHEF1/QuEst++-AROW	27.36
RTM-DCU/s5-RTM-GLMd	25.92
SHEF1/QuEst++-PA	25.49
RTM-DCU/s4-RTM-GLMd	24.95
Baseline	0.2044

Table 13: Alternative results for the WMT15 quality estimation Task 2 according to the sequence correlation metric. The winning submissions are indicated by a •. These are the top-scoring submission and those that are not significantly worse according to approximate randomization tests with 95% confidence intervals. Submissions whose results are statistically different from others according to the same test are grouped by a horizontal line.

The sequence correlation metric gives preference to systems that use sequence labelling (modelling dependencies between the assigned tags). We consider this a desirable feature, as we are generally not interested in maximising the prediction accuracy for individual words, but in maximising the accuracy for word-level labelling in the context of the whole sentence. However, using the TER alignment to tag errors cannot capture "phrase-level errors", and each token is considered independently when the dataset is built. This is a fundamental issue with the current definition of the word-level quality estimation that we intend to address in future work.

Our intuition is that the sequence correlation metric should be closer to human perception of word-level QE than F_1 scores. The goal of word-level QE is to identify incorrect segments of a sentence — and the sequence correlation metric evaluates how good the segmentation of the sentence is into correct and incorrect phrases. A system can get very high F_1 score by (almost) randomly assigning a correct tag to a word, and giving very little information on correct and incorrect areas in the text. That was illustrated by the WMT14 word-level QE task results, where the baseline strategy that assigned tag "BAD" to all words had significantly higher F_1 score than any of the submissions. fundamental problem with the current task. I added a sentence about it at the end of the paragraph before this one.

4.5 Task 3: Predicting document-level quality

Predicting the quality of units larger than sentences can be useful in many scenarios. For example, consider a user searching for information about a product on the web. The user can only find reviews in German but he/she does not speak the language, so he/she uses an MT system to translate the reviews into English. In this case, predictions on the quality of individual sentences in a translated review are not as informative as predictions on the quality of the entire review.

With the goal of exploring quality estimation beyond sentence level, this year we proposed a document-level task for the first time. Due to the lack of large datasets with machine translated documents (by various MT systems), we consider short paragraphs as *documents*. The task consisted in scoring and ranking paragraphs according to their predicted quality.

Data The paragraphs were extracted from the WMT13 translation task test data (Bojar et al., 2013), using submissions from all participating MT systems. Source paragraphs were randomly chosen using the paragraph markup in the SGML files. For each source paragraph, a translation was taken from a different MT system such as to select approximately the same number of instances from each MT system. We considered EN-DE and DE-EN as language pairs, extracting $1,215$ paragraphs for each language pair. 800 paragraphs were used for training and 415 for test.

Since no human annotation exists for the quality of entire paragraphs (or documents), Meteor against reference translations was used as quality label for this task. Meteor was calculated using its implementation within the Asyia toolkit, with the following settings: exact match, tokenised and case insensitive (Giménez and Màrquez, 2010).

Evaluation The evaluation of the paragraph-level task was the same as that for the sentence-level task. MAE and RMSE are reported as evaluation metrics for the scoring task, with MAE as official metric for systems ranking. For the ranking task, DeltaAvg and Spearman's ρ correlation are reported, with DeltaAvg as official metric for systems ranking. To evaluate the significance of the results, bootstrap resampling (1K times) with 95% confidence intervals was used. Pearson's r correlation scores with the Williams significance test are also reported.

Results Table 14 summarises the results of the ranking variant of Task 3.[21] They are sorted from best to worst using the DeltaAvg metric scores as primary key and the Spearman's ρ rank correlation scores as secondary key. RTM-DCU submissions achieved the best scores: RTM-SVR was the winner for EN-DE, and RTM-FS-SVR for DE-EN. For EN-DE, the HIDDEN system did not show significant difference against the baseline. For DE-EN, USHEF/QUEST-DISC-BO, USAAR-USHEF/BFF and HIDDEN were not significantly different from the baseline.

The results of the scoring variant are given in Table 15, sorted from best to worst by using the MAE metric scores as primary key and the RMSE metric scores as secondary key. Again the RTM-DCU submissions scored the best for both language pairs. All systems were significantly better than the baseline. However, the difference between the baseline system and all submissions was much lower in the scoring evaluation than in the ranking evaluation.

Following the suggestion in (Graham, 2015), Table 16 shows an alternative ranking of systems considering Pearson's r correlation results. The alternative ranking differs from the official ranking in terms of MAE: for EN-DE, RTM-DCU/RTM-FS-SVR is no longer in the winning group, while for DE-EN, USHEF/QUEST-DISC-BO and USAAR-USHEF/BFF did not show statistically significant difference against the baseline. However, as with Task 1 these results are the same as the official ones in terms of DeltaAvg.

4.6 Discussion

In what follows, we discuss the main findings of this year's shared task based on the goals we had previously identified for it.

Advances in sentence- and word-level QE

For sentence-level prediction, we used similar data and quality labels as in previous editions of the task: English-Spanish, news text domain and HTER labels to indicate post-editing effort. The main differences this year were: (i) the much larger size of the dataset, (ii) the way post-editing was performed – by a large number of crowd-sourced translators, and (iii) the MT systems used – an online statistical system. We will discuss items (i) and (ii) later in this section. Regarding (iii), the main implication of using an online system was that one could not have access to many of the resources commonly used to extract features, such as the SMT training data and lexical tables. As a consequence, surrogate resources were used for certain features, including many of the baseline ones, which made them less effective. To avoid relying on such resources, novel features were explored, for example those based on deep neural network architectures (word embeddings and continuous space language models by SHEFF-NN) and those based on pseudo-references (n-gram overlap and agreement features by LORIA).

While it is not possible to compare results directly with those published in previous years, for sentence level we can observe the following with respect to the corresponding task in WMT14 (Task 1.2):

[21] Results for MAE, RMSE and DeltaAvg are multiplied by 100 to improve readability.

System ID	**DeltaAvg** ↑	Spearman's ρ ↑
English-German		
• RTM-DCU/RTM-SVR	7.62	−0.62
RTM-DCU/RTM-FS-SVR	6.45	−0.67
USHEF/QUEST-DISC-REP	4.55	0.32
USAAR-USHEF/BFF	3.98	0.27
Baseline SVM	1.60	0.14
HIDDEN	1.04	0.05
German-English		
• RTM-DCU/RTM-FS-SVR	4.93	−0.64
RTM-DCU/RTM-FS+PLS-SVR	4.23	−0.55
USHEF/QUEST-DISC-BO	1.55	0.19
Baseline SVM	0.59	0.05
USAAR-USHEF/BFF	0.40	0.12
HIDDEN	0.12	−0.03

Table 14: Official results for the ranking variant of the WMT15 quality estimation Task 3. The winning submissions are indicated by a •. These are the top-scoring submission and those that are not significantly worse according to bootstrap resampling (1K times) with 95% confidence intervals. The systems in the gray area are not different from the baseline system at a statistically significant level according to the same test.

System ID	**MAE** ↓	RMSE ↓
English-German		
• RTM-DCU/RTM-FS-SVR	7.28	11.96
• RTM-DCU/RTM-SVR	7.5	11.35
USAAR-USHEF/BFF	9.37	13.53
USHEF/QUEST-DISC-REP	9.55	13.46
Baseline SVM	10.05	14.25
German-English		
• RTM-DCU/RTM-FS-SVR	4.94	8.74
RTM-DCU/RTM-FS+PLS-SVR	5.78	10.70
USHEF/QUEST-DISC-BO	6.54	10.10
USAAR-USHEF/BFF	6.56	10.12
Baseline SVM	7.35	11.40

Table 15: Official results for the scoring variant of the WMT15 quality estimation Task 3. The winning submissions are indicated by a •. These are the top-scoring submission and those that are not significantly worse according to bootstrap resampling (1K times) with 95% confidence intervals. The systems in the gray area are not different from the baseline system at a statistically significant level according to the same test.

- In terms of scoring, according to the primary metric – MAE, in WMT15 all systems except one were significantly better than the baseline. In both WMT14 and WMT15 only one system was significantly worse than the baseline. However, in WMT14 four others (out of nine) performed no different than the baseline. This year, no system tied with the baseline: the remaining seven systems were significantly better than the baseline. One could say systems are consistently better this year. It is worth mentioning that the baseline remains the same, but as previously noted, the resources used to extract baseline features are likely to be less useful this year given the mismatch between the data used to produce them and the data used to build the online SMT system.

- In terms of ranking, in WMT14 one system was significantly worse than the baseline, and the four remaining systems were significantly better. This year, all eight submissions are significantly better than the baseline. This can once more be seen as progress from last year's results. These results as well as the general ranking of systems were also found following Pearson's correlation as metric, as

System ID	**Pearson's r ↑**
English-German	
• RTM-DCU/RTM-SVR	0.59
RTM-DCU/RTM-FS-SVR	0.53
USHEF/QUEST-DISC-REP	0.30
USAAR-USHEF/BFF	0.29
Baseline SVM	0.12
German-English	
• RTM-DCU/RTM-FS-SVR	0.52
RTM-DCU/RTM-FS+PLS-SVR	0.39
USHEF/QUEST-DISC-BO	0.10
USAAR-USHEF/BFF	0.08
Baseline SVM	0.06

Table 16: Alternative results for the scoring variant of the WMT15 quality estimation Task 3. The winning submissions are indicated by a •. These are the top-scoring submission and those that are not significantly worse according to the Williams test with 95% confidence intervals. The systems in the gray area are not different from the baseline system at a statistically significant level according to the same test.

suggested by Graham (2015).

For the word level task, a comparison with the WMT14 corresponding task is difficult to perform, as in WMT14 we did not have a meaningful baseline. The baseline used then for binary classification was to tag all words with the label "BAD". This baseline outperformed all the submissions in terms of F_1 for the "BAD" class, but it cannot be considered an appropriate baseline strategy (see Section 4.4). This year the submissions were compared against the output of a real baseline system and the set of baseline features was made available to participants. Although the baseline system itself performed worse than all the submitted systems, some other systems benefited from adding baseline features to their feature sets (UAlacant, UGENT, HDCL).

Considering the feature sets and methods used, the number of participants in the WMT14 word-level task was too small to draw reliable conclusion: four systems for English–Spanish and one system for all other three language pairs. The larger number of submissions this year is already a positive result: 16 submissions from eight teams. Inspecting the systems submitted this and last year, we can speculate about the most promising techniques. Last year's winning system used a neural network trained on pseudo-reference features (namely, features extracted from n-best lists) (Camargo de Souza et al., 2014). This year's winning systems are also based on pseudo-reference features (UAlacant) and deep neural network architectures (HDCL). Luong et al. (2013) had previously reported that pseudo-reference features improve the accuracy of word-level predictions. The two most recent editions of this shared task seem to indicate that the state of the art in word-level quality estimation relies upon such features, as well as the ability to model the relationship between the source and target languages using large datasets.

Effectiveness of quality labels, features and learning methods for document-level QE

The task of paragraph-level prediction received fewer submissions than the other two tasks: four submissions for the scoring variant and five for the ranking variant, for both language pairs. This is understandable as it was the first time the task was run. Additionally, paragraph-level QE is still fairly new as a task. However, we were able to draw some conclusions and learn valuable lessons for future research in the area.

By and large, most features are similar to those used for sentence-level prediction. Discourse-aware features showed only marginal improvements relative to the baseline system (USHEF systems for EN-DE and DE-EN). One possible reason for that is the way the training and test data sets were created, including paragraphs with only one sentence. Therefore, discourse features could not be fully explored as they aim to model relationships and dependencies across sentences, as well as within sentences. In future, data will be selected more carefully in order to consider only paragraphs or documents with more sentences.

Systems applying feature selection techniques, such as USAAR-USHEF/BFF, did not obtain major improvements over the baseline. However, they provided interesting insights by finding a minimum set of baseline features that can be used to build models with the same performance as the entire baseline feature set. These are models with only three features selected as the best combination by exhaustive search.

The winning submissions for both language pairs and variants – RTM-DCU – explored features based on the source and target side information. These include distributional similarity, closeness of test instances to the training data, and indicators for translation quality. External data was used to select "interpretants", which contain data close to both training and test sets to provide context for similarity judgements.

In terms of quality labels, one problem observed in previous work on document-level QE (Scarton et al., 2015b) is the low variation of scores (in this case, Meteor) across instances of the dataset. Since the data collected for this task included translations from many different MT systems, this was not the case. Table 17 shows the average and standard deviation (STDEV) values for the datasets (both training and test set together). Although the variation is substantial, the average value of the training set is a good predictor. In other words, if we consider the average of the training set scores as the prediction value for all data points in the test set, we obtain results as good as the baseline system. For our datasets, the MAE figure for EN-DE is 10, and for DE-EN 7 – the same as the baseline system. We can only speculate that automatically assigned quality labels based on reference translations such as Meteor are not adequate for this task. Other automatic metrics tend to behave similarly to Meteor for document-level (Scarton et al., 2015b). Therefore, finding an adequate quality label for document-level QE remains an open issue. Having humans directly assign quality labels is much more complex than in the sentence and word level cases. Annotation of entire documents, or even paragraphs, becomes a harder, more subjective and much more costly task. For future editions of this task, we intend to collect datasets with human-targeted document-level labels obtained indirectly, e.g. through post-editing.

No submission focused on exploring learning algorithms specifically targeted at document-level prediction.

	EN-DE		DE-EN	
	AVG	STDEV	AVG	STDEV
Meteor (↑)	0.35	0.14	0.26	0.09

Table 17: Average metric scores for automatic metrics in the corpus for Task 3.

Differences between sentence-level and document-level QE

The differences between sentence and document-level prediction have not been explored to a great extent. Apart from the discourse-aware features by USHEF, the baseline and other features explored by participating teams for document level prediction were simple aggregations of sentence level feature values.

Also, none of the submitted systems use sentence-level predictions as features for paragraph-level QE. Although this technique is possible in principle, its effectiveness has not yet been proved. (Specia et al., 2015) report promising results when using word-level prediction for sentence-level QE, but inclusive results when using sentence-level prediction for document-level QE. They considered BLEU, TER and Meteor as quality labels, all leading to similar findings. Once more the use of inadequate quality labels for document-level prediction could have been the reason.

No submission evaluated different machine learning algorithms for this task. The same algorithms as those used for sentence-level prediction were applied by all participating teams.

Effect of training data sizes and quality for sentence and word-level QE

As it was previously mentioned, the post-editions used for this year's sentence and word-level tasks were obtained through a crowdsourcing platform where translators volunteered to post-edit machine translations. As such, one can expect that not all post-editions will reach the highest standards of professional translation. Manual inspection of a small sample of the data, however, showed that the post-editions were high quality, although stylistic differences are evident in some cases. This is likely due to the fact that different editors, with different styles and levels of expertise, worked on different segments. Another factor that may have influenced the quality of the post-editions is the

fact that segments were fixed out of context. For word level, in particular, a potential issue is the fact that the labelling of the words was done completely automatically, using a tool for alignment based on minimum edit distance (TER).

On the positive side, this dataset is much larger dataset than any we have used before for prediction at any level: nearly 12K segments for training/development, as opposed to maximum 2K in previous years. For sentence-level prediction we did not expect massive gains from larger datasets, as it has been shown that small amounts of data can be as effective or even more effective than the entire collection, if selected in a clever way (Beck et al., 2013a,b). However, it is well known that data sparsity is an issue for word-level prediction, so we expected a large dataset to improve results considerably for this task.

Unfortunately, having access to a large number of samples did not seem to bring much improvement for word-level predictions accuracy. The main reason for that was the fact that the number of erroneous words in the training data was too small, as compared to the number of correct words: 50% of the sentences had zero incorrect words (15% of the sentences) or fewer than 15% incorrect words (35% of the sentences). Participants used various data manipulation strategies to improve results: filtering of the training data, as in DCU-SHEFF systems, alternative labelling of the data which discriminates between "OK" label in the beginning, middle, and end of a good segment, and insertion of additional incorrect words, as in SAU-KERC submissions. Additionally, most participants in the word-level task leveraged additional data in some way, which points to the need for even larger but more varied post-edited datasets in order to make significant progress in this task.

5 Automatic Post-editing Task

This year WMT hosted for the first time a shared task on automatic post-editing (APE) for machine translation. The task requires to automatically correct the errors present in a machine translated text. As pointed out in Parton et al. (2012) and Chatterjee et al. (2015b), from the application point of view, APE components would make it possible to:

- Improve MT output by exploiting information unavailable to the decoder, or by performing deeper text analysis that is too expensive at the decoding stage;
- Cope with systematic errors of an MT system whose decoding process is not accessible;
- Provide professional translators with improved MT output quality to reduce (human) post-editing effort;
- Adapt the output of a general-purpose MT system to the lexicon/style requested in a specific application domain.

The first pilot round of the APE task focused on the challenges posed by the "black-box" scenario in which the MT system is unknown and cannot be modified. In this scenario, APE methods have to operate at the downstream level (that is *after* MT decoding), by applying either rule-based techniques or statistical approaches that exploit knowledge acquired from human post-editions provided as training material. The objectives of this pilot were to: *i)* define a sound evaluation framework for the task, *ii)* identify and understand the most critical aspects in terms of data acquisition and system evaluation, *iii)* make an inventory of current approaches and evaluate the state of the art and *iv)* provide a milestone for future studies on the problem.

5.1 Task description

Participants were provided with training and development data consisting of (*source*, *target*, *human post-edition*) triplets, and were asked to return automatic post-editions for a test set of unseen (*source*, *target*) pairs.

Data

Training, development and test data were created by randomly sampling from a collection of English-Spanish (*source*, *target*, *human post-edition*) triplets drawn from the `news` domain.[22] Instances were sampled after applying a series of data cleaning steps aimed at removing duplicates and those triplets in which any of the elements (source, target, post-edition) was either too long or too short compared to the others, or included tags or special problematic symbols. The main reason for random sampling was to induce some homogeneity across the three datasets and, in turn,

[22] The original triplets were provided by Unbabel (`https://unbabel.com/`).

to increase the chances that correction patterns learned from the training set can be applied also to the test set. The downside of losing information yielded by text coherence (an aspect that some APE systems might take into consideration) has hence been accepted in exchange for a higher error repetitiveness across the three datasets. Table 18 provides some basic statistics about the data.

The training and development sets respectively consist of $11,272$ and $1,000$ instances. In each instance:

- The source (SRC) is a tokenized English sentence having a length of at least 4 tokens. This constraint on the source length was posed in order to increase the chances to work with grammatically correct full sentences instead of phrases or short keyword lists;
- The target (TGT) is a tokenized Spanish translation of the source, produced by an unknown MT system;
- The human post-edition (PE) is a manually-revised version of the target. PEs were collected by means of a crowdsourcing platform developed by the data provider.

Test data ($1,817$ instances) consists of (*source*, *target*) pairs having similar characteristics of those in the training set. Human post-editions of the test target instances were left apart to measure system performance.

The data creation procedure adopted, as well as the origin and the domain of the texts pose specific challenges to the participating systems. As discussed in Section 5.4, the results of this pilot task can be partially explained in light of such challenges. This dataset, however, has three major advantages that made it suitable for the first APE pilot: *i)* it is relatively large (hence suitable to apply statistical methods), *ii)* it was not previously published (hence usable for a fair evaluation), *iii)* it is freely available (hence easy to distribute and use for evaluation purposes).

Evaluation metric

System performance is evaluated by computing the distance between *automatic* and *human* post-editions of the machine-translated sentences present in the test set (*i.e.* for each of the *1,817* target test sentences). This distance is measured in terms of Translation Error Rate (TER) (Snover et al., 2006a), an evaluation metric commonly used in MT-related tasks (*e.g.* in quality estimation) to measure the minimum edit distance between an automatic translation and a reference translation.[23] Systems are ranked based on the average TER calculated on the test set by using the TERcom[24] software: lower average TER scores correspond to higher ranks. Each run is evaluated in two modes, namely: *i)* case insensitive and *ii)* case sensitive. Evaluation scripts to compute TER scores in both modalities have been made available to participants through the APE task web page.[25]

Baseline

The official baseline is calculated by averaging the distances computed between the raw MT output and the human post-edits. In practice, the baseline APE system is a system that leaves all the test targets unmodified.[26] Baseline results computed for both evaluation modalities (case sensitive/insensitive) are reported in Tables 20 and 21.

Monolingual translation as another term of comparison. To get further insights about the progress with respect to previous APE methods, participants' results are also analysed with respect to another term of comparison: a re-implementation of the state-of-the-art approach firstly proposed by Simard et al. (2007).[27] For this purpose, a phrase-based SMT system based on Moses (Koehn et al., 2007) is used. Translation and reordering models were estimated following the Moses protocol with default setup using MGIZA++ (Gao and Vogel, 2008) for word alignment. For language modeling we used the

[23] Edit distance is calculated as the number of edits (word insertions, deletions, substitutions, and shifts) divided by the number of words in the reference. Lower TER values indicate better MT quality.

[24] `http://www.cs.umd.edu/~snover/tercom/`

[25] `http://www.statmt.org/wmt15/ape-task.html`

[26] In this case, since edit distance is computed between each machine-translated sentence and its human-revised version, the actual evaluation metric is the human-targeted TER (HTER). For the sake of clarity, since TER and HTER compute edit distance in the same way (the only difference is in the origin of correct sentence used for comparison), henceforth we will use TER to refer to both metrics.

[27] This is done based on the description provided in Simard et al. (2007). Our re-implementation, however, is not meant to officially represent such approach. Discrepancies with the actual method are indeed possible due to our misinterpretation or to wrong guesses about details that are missing in the paper.

	Tokens			Types			Lemmas		
	SRC	TGT	PE	SRC	TGT	PE	SRC	TGT	PE
Train (11,272)	238,335	257,643	257,879	23,608	25,121	27,101	13,701	7,624	7,689
Dev (1,000)	21,617	23,213	23,098	5,482	5,760	5,966	3,765	2,810	2,819
Test (1,817)	38,244	40,925	40,903	7,990	8,498	8,816	5,307	3,778	3,814

Table 18: Data statistics.

KenLM toolkit (Heafield, 2011) for standard n-gram modeling with an n-gram length of 5. Finally, the APE system was tuned on the development set, optimizing TER with Minimum Error Rate Training (Och, 2003). The results of this additional term of comparison, computed for both evaluation modalities (case sensitive/insensitive), are also reported in Tables 20 and 21.

For each submitted run, the statistical significance of performance differences with respect to the baseline and the re-implementation of Simard et al. (2007) is calculated with the bootstrap test (Koehn, 2004).

5.2 Participants

Four teams participated in the APE pilot task by submitting a total of seven runs. Participants are listed in Table 19; a short description of their systems is provided in the following.

Abu-MaTran. The Abu-MaTran team submitted the output of two statistical post-editing (SPE) systems, both relying on the MOSES phrase-based statistical machine translation toolkit (Koehn et al., 2007) and on sentence level classifiers. The first element of the pipeline, the SPE system, is trained on the automatic translation of the News Commentary v8 corpus from English to Spanish aligned with its reference. This translation is obtained with an out-of-the-box phrase-based SMT system trained on Europarl v7. Both translation and post-editing systems use a 5-gram Spanish LM with modified Kneser-Ney smoothed trained on News Crawl 2011 and 2012 with KenLM (Heafield, 2011). For the second element of the pipeline, a binary classifier to select the best translation between the given MT output or its automatic post-edition is used. Two different approaches are investigated: a 180-hand-crafted-based regression model trained with a Support Vector Machine (SVM) with a radial basis function kernel to estimate the sentence-level HTER score, and a Recurrent Neural Network (RNN) classifier using context word embeddings as input and classifying each word of a sentence as *good* or *bad*. An automatic translation to be post-edited is first decoded by our SPE system, then fed into one of the classifiers identified as SVM180feat and RNN. The HTER estimator selects the translation with the lower score while the binary word-level classifier selects the translation with the fewer amount of *bad* tags. The official evaluation of the shared task show an advantage of the RNN approach compared to SVM.

FBK. The two runs submitted by FBK (Chatterjee et al., 2015a) are based on combining the statistical phrase-based post-editing approach proposed by Simard et al. (2007) and its most significant variant proposed by Béchara et al. (2011). The APE systems are built-in an incremental manner. At each stage of the APE pipeline, the best configuration of a component is decided and then used in the next stage. The APE pipeline begins with the selection of the best language model from several language models trained on different types and quantities of data. The next stage addresses the possible data sparsity issues raised by the relatively small size of the training data. Indeed, an analysis of the original phrase table obtained from the training set revealed that a large part of its entries is composed of instances that occur only once in the training. This has the obvious effect of collecting potentially unreliable "translation" (or, in the case of APE, *correction*) rules. The problem is exacerbated by the "context-aware" approach proposed by Béchara et al. (2011), which builds the phrase table by joining source and target tokens thus breaking down the co-occurrence counts into smaller numbers. To cope with this problem, a novel feature (*neg-impact*) is designed to prune the phrase table by measuring the usefulness of each translation. The higher is the value of the *neg-impact* feature, the less useful is the translation option. After pruning, the final stage of the APE pipeline tries to raise the capability of the decoder to select the correct translation rule by the introduction of new task specific features integrated in

ID	Participating team
Abu-MaTran	Abu-MaTran Project (Prompsit)
FBK	Fondazione Bruno Kessler, Italy (Chatterjee et al., 2015a)
LIMSI	Laboratoire d'Informatique pour la Mécanique et les Sciences de l'Ingénieur, France (Wisniewski et al., 2015)
USAAR-SAPE	Saarland University, Germany & Jadavpur University, India (Pal et al., 2015b)

Table 19: Participants in the WMT15 Automatic Post-editing pilot task.

the model. These features measure the similarity and the reliability of the translation options and help to improve the precision of the resulting APE system.

LIMSI. For the first edition of the APE shared task LIMSI submitted two systems (Wisniewski et al., 2015). The first one is based on the approach of Simard et al. (2007) and considers the APE task as a monolingual translation between a translation hypothesis and its post-edition. This straightforward approach does not succeed in improving translation quality. The second submitted system implements a series of sieves, each applying a simple post-editing rule. The definition of these rules is based on an analysis of the most frequent error corrections and aims at: *i*) predicting word case; *ii*) predicting exclamation and interrogation marks; and *iii*) predicting verbal endings. Experiments with this approach show that this system also hurts translation quality. An in-depth analysis revealed that this negative result is mainly explained by two reasons: *i*) most of the post-edition operations are nearly unique, which makes very difficult to generalize from a small amount of data; and *ii*) even when they are not, the high variability of post-editing, already pointed out by Wisniewski et al. (2013), results in predicting legitimate corrections that have not been made by the annotators, therefore preventing from improving over the baseline.

USAAR-SAPE. The USAAR-SAPE system (Pal et al., 2015b) is designed with three basic components: corpus preprocessing, hybrid word alignment and a state-of-the-art phrase-based SMT system integrated with the hybrid word alignment. The preprocessing of the training corpus is carried out by stemming the Spanish MT output and the PE data using Freeling (Padr and Stanilovsky, 2012). The hybrid word alignment method combines different kinds of word alignment: GIZA++ word alignment with the grow-diag-final-and (GDFA) heuristic (Koehn, 2010), SymGiza++ (Junczys-Dowmunt and Szal, 2011), the Berkeley aligner (Liang et al., 2006), and the edit distance-based aligners (Snover et al., 2006a; Lavie and Agarwal, 2007). These different word alignment tables (Pal et al., 2013) are combined by a mathematical union method. For the phrase-based SMT system various maximum phrase lengths for the translation model and n–gram settings for the language model are used. The best results in terms of BLEU (Papineni et al., 2002) score are achieved by a maximum phrase length of 7 for the translation model and a 5-gram language model.

5.3 Results

The official results achieved by the participating systems are reported in Tables 20 and 21. The seven runs submitted are sorted based on the average TER they achieve on test data. Table 20 shows the results computed in case sensitive mode, while Table 21 provides scores computed in the case insensitive mode.

Both rankings reveal an unexpected outcome: none of the submitted runs was able to beat the baselines (*i.e.* average TER scores of 22.91 and 22.22 respectively for case sensitive and case insensitive modes). All differences with respect to such baselines, moreover, are statistically significant. In practice, this means that what the systems learned from the available data was not reliable enough to yield valid corrections of the test instances. A deeper discussion about the possible causes of this unexpected outcome is provided in Section 5.4.

Unsurprisingly, for all participants the case insensitive evaluation results are slightly better than the case sensitive ones. Although the two rankings are not identical, none of the systems was particularly penalized by the case sensitive evaluation. Indeed, individual differences in the two modes are always close to the same value (~ 0.7 TER difference) measured for the two baselines.

ID	Avg. TER
Baseline	22.913
FBK Primary	23.228
LIMSI Primary	23.331
USAAR-SAPE	23.426
LIMSI Contrastive	23.573
Abu-MaTran Primary	23.639
FBK Contrastive	23.649
(Simard et al., 2007)	23.839
Abu-MaTran Contrastive	24.715

Table 20: Official results for the WMT15 Automatic Post-editing task – average TER (↓) case sensitive.

ID	Avg. TER
Baseline	22.221
LIMSI Primary	22.544
FBK Primary	22.551
USAAR-SAPE	22.710
Abu-MaTran Primary	22.769
LIMSI Contrastive	22.861
FBK Contrastive	22.949
(Simard et al., 2007)	23.130
Abu-MaTran Contrastive	23.705

Table 21: Official results for the WMT15 Automatic Post-editing task – average TER (↓) case insensitive.

In light of this, and considering the importance of case sensitive evaluation in some language settings (*e.g.* having German as target), future rounds of the task will likely prioritize this more strict evaluation mode.

Overall, the close results achieved by participants reflect the fact that, despite some small variations, all systems share the same underlying statistical approach of Simard et al. (2007). As anticipated in Section 5.1, in order to get a rough idea about the extent of the improvements over such state-of-the-art method, we replicated it and considered its results as another term of comparison in addition to the baselines. As shown in Tables 20 and 21, the performance results achieved by our implementation of Simard et al. (2007) are 23.839 and 23.130 in terms of TER for the respective case sensitive and insensitive evaluations. Compared to these scores, most of the submitted runs achieve better performance, with positive average TER differences that are always statistically significant. We interpret this as a good sign: despite the difficulty of the task, the novelties introduced by each system allowed to make significant steps forward with respect to a prior reference technique.

5.4 Discussion

To better understand the results and gain useful insights about this pilot evaluation round, we perform two types of analysis. The first one is focused on the **data**, and aims to understand the possible reasons of the difficulty of the task. In particular, by analysing the challenges posed by the *origin* and the *domain* of the text material used, we try to find indications for future rounds of the APE task. The second type of analysis focuses on the **systems** and their behaviour. Although they share the same underlying approach and achieve similar results, we aim to check if interesting differences can be captured by a more fine grained analysis that goes beyond rough TER measurements.

Data analysis

In this section we investigate the possible relation between participants' results and the nature of the data used in this pilot task (*e.g.* quantity, sparsity, domain and origin) . For this purpose, we take advantage of a new dataset – the Autodesk Post-Editing Data corpus[28] – which has been publicly released after the organisation of the APE pilot task. Although it was not usable for this first round, its characteristics make it particularly suitable for our analysis purposes. In particular: *i)* Autodesk data predominantly covers the domain of `software user manuals` (that is, a restricted domain compared to a general one like `news`), and *ii)* post-edits come from professional translators (that is, at least in principle, a more reliable source of corrections compared to crowd-sourced workforce). To guarantee a fair comparison, English-Spanish (*source*, *target*, *human post-edition*) triplets drawn from the Autodesk corpus are split in training, development and test sets under the constraint that the total number of target words and the TER in each set should be similar to the APE task splits. In this setting, performance differences between systems trained on the two datasets will only depend on the different nature of the data (*e.g.* domain). Statistics of the training sets are reported in Table 22 (those concerning the

[28]The corpus (`https://autodesk.app.box.com/Autodesk-PostEditing`) consists of parallel English source-MT/TM target segments post-edited into several languages (Chinese, Czech, French, German, Hungarian, Italian, Japanese, Korean, Polish, Brazilian Portuguese, Russian, *Spanish*) with between 30K to 410K segments per language.

		APE Task	Autodesk
Tokens	SRC	238,335	220,671
	TGT	257,643	257.380
	PE	257,879	260,324
Types	SRC	23,608	11,858
	TGT	25,121	11,721
	PE	27,101	12,399
Lemmas	SRC	13,701	5,092
	TGT	7,624	3,186
	PE	7,689	3,334
RR	SRC	2.905	6.346
	TGT	3.312	8.390
	PE	3.085	8.482

Table 22: WMT APe Task and Autodesk training data statistics.

APE task data are the same of Table 18).

The impact of data sparsity. A key issue in most evaluation settings is the representativeness of the training data with respect to the test set used. In the case of the statistical approach at the core of all the APE task submissions, this issue is even more relevant given the limited amount of training data available. In the APE scenario, data representativeness relates to the fact that the correction patterns learned from the training set can be applied also to the test set (as mentioned in Section 5.1, in the data creation phase random sampling from an original data collection was applied for this purpose). From this point of view, dealing with restricted domains such as `software user manuals` should be easier than working with `news` data. Indeed, restricted domains are more likely to feature smaller vocabularies, be more repetitive (or, in other terms, less sparse) and, in turn, determine a higher applicability of the learned error correction patterns.

To check the relation between task difficulty and data repetitiveness, we compared different monolingual indicators (*i.e.* number of types and lemmas, and repetition rate[29] – RR) computed on the APE and the Autodesk source, target and post-edited sentences. Although both the datasets have the same amount of target tokens, Table 22 shows that the APE training set has nearly double of types and lemmas compared to the Autodesk data, which indicates the presence of less repeated information. A similar conclusion can be drawn by observing that the Autodesk dataset has a repetition rate that is more than twice the value computed for the APE task data.

This monolingual analysis does not provide any information about the level of repetitiveness of the correction patterns made by the post-editors, because it does not link the target and the post-edited sentences. To investigate this aspect, two instances of the re-implemented approach of Simard et al. (2007) introduced in Section 5.1 are respectively trained on the APE and the Autodesk training sets. We consider the distribution of the frequency of the translation options in the phrase table as a good indicator of the level of repetitiveness of the corrections in the data. For instance, a large number of translation options that appear just one or only few times in the data indicates a higher level of sparseness. As expected due to the limited size of the training set, the vast majority of the translation options in both phrase tables are singletons as shown in Table 23. Nevertheless, the Autodesk phrase table is more compact (731k versus 1,066k) and contains 10% fewer singletons than the APE task phrase table. This confirms that the APE task data is more sparse and suggests that it might be easier to learn more applicable correction patterns from the Autodesk domain-specific data.

To verify this last statement, the two APE systems are evaluated on their own test sets. As previously shown, the system trained on the APE task data is not able to improve over the performance achieved by a system that leaves all the test targets unmodified (see Table 20). On the contrary, starting from a baseline of 23.57, the system trained on the Autodesk data is able to reduce the TER by 3.55 points (20.02). Interestingly, the Autodesk APE system is able to correctly fix the target sentences and improve the TER by 1.43 points even with only 25% of the training data. These results confirm our intuitions about the usefulness of repetitive data and show that, at least in restricted-domain scenarios, automatic post-editing can be successfully used as an aid to improve the output of an MT system.

Professional vs. Crowdsourced post-editions Differently from the Autodesk data, for which the post-editions are created by professional translators, the APE task data contains crowdsourced MT corrections collected from unknown (likely non-

[29]Repetition rate measures the repetitiveness inside a text by looking at the rate of non-singleton n-gram types (n=1. . .4) and combining them using the geometric mean. Larger value means more repetitions in the text. For more details see Cettolo et al. (2014)

	Percentage of Phrase Pairs	
Phrase Pair Count	APE 2015 Training	Autodesk
1	95.2%	84.6%
2	2.5%	8.8%
3	0.7%	2.7%
4	0.3%	1.2%
5	0.2%	0.6%
6	0.15%	0.4%
7	0.10%	0.3%
8	0.07%	0.2%
9	0.06%	0.2%
10	0.04%	0.1%
> 10	0.3%	0.9%
Total Entries	1,066,344	703,944

Table 23: Phrase pair count distribution in two phrase tables built using the APE 2015 training and the Autodesk dataset.

expert) translators. One risk, given the high variability of valid MT corrections, is that the crowdsourced workforce follows post-editing attitudes and criteria that differ from those of professional translators. Professionals tend to: *i)* maximize productivity by doing only the necessary and sufficient corrections to improve translation quality, and *ii)* follow consistent translation criteria, especially for domain terminology. Such a tendency will likely result in coherent and minimally post-edited data from which learning and drawing statistics is easier. This is not guaranteed by crowdsourced workers which do not have specific time or consistency constraints. This suggests that non-professional post-editions and the correction patterns learned from them will feature less coherence, higher noise and higher sparsity.

To assess the potential impact of these issues on data representativeness (and, in turn, on the task difficulty), we analyse a subset of the APE test instances (221 triples randomly sampled) in which target sentences were post-edited by professional translators. The analysis focuses on TER scores computed between:

1. The target sentences and their crowdsourced post-editions (avg. TER = 26.02);

2. The target sentences and their professional post-editions (avg. TER = 23.85);

3. The crowdsourced post-editions and the professional ones, using the latter as references (avg. TER = 29.18).

The measured values indicate an attitude of non-professionals to correct *more often* and *differently* from the professional translators. Interestingly, and similar to the findings of Potet et al. (2012), crowdsourced post-editions feature a distance from the professional ones that is even higher than the distance between the original target sentences and the experts' corrections (29.18 vs. 23.85). If we consider the output of professional translators as a gold standard (made of coherent and minimally post-edited data), these figures suggest a higher difficulty in handling crowdsourced corrections.

Further insights can be drawn from the analysis of the word level corrections produced by the two translator profiles. To this aim, word insertions, deletions, substitutions and phrase shifts are extracted using the TERcom software similar to Blain et al. (2012) and Wisniewski et al. (2013). For each error type, the ratio between the number of edit operations and the total number of occurred errors operations performed is computed. This quantity provides us with a measure of the level of repetitiveness of the errors, with 100% indicating that all the error patterns are unique, and small values indicating that most of the errors are repeated. Our results show that non-experts have generally larger ratio values than the professional translators (insertion +6%, substitution +4%, deletion +4%). This seems to support our hypothesis that, independently from their quality, post-editions collected from non-experts are less coherent than those derived from professionals. It is unlikely that different crowdsourced workers will apply the same corrections in the same contexts. If this hypothesis holds, the difficulty of this APE pilot task could be partially ascribed to this unavoidable intrinsic property of crowdsourced data. This aspect, however, should be further investigated to draw definite conclusions.

System/performance analysis

The TER results presented in Tables 20 and 21 evidence small differences between participants, but they do not shed light on the real behaviour of the systems. To this aim, in this section the submitted runs are analysed by taking into consideration the changes made by each system to the test instances (case sensitive evaluation mode). In particular, Table 24 provides the number of modified, improved and deteriorated sentences, together with the percentage of edit operations performed (insertions,

ID	Modified Sentences	Improved Sentences	Deteriorated Sentences	Edit operations			
				Ins	Del	Sub	Shifts
FBK Primary	276	64	147	17.8	17.8	55.9	8.5
LIMSI Primary	339	75	217	19.4	16.8	55.2	8.6
USAAR-SAPE	422	53	229	17.6	17.4	56.7	8.4
LIMSI Contrastive	454	61	260	17.4	19.0	55.3	8.3
Abu-MaTran Primary	275	8	200	17.7	17.2	56.8	8.2
FBK Contrastive	422	52	254	18.4	17.0	56.2	8.4
Abu-MaTran Contrastive	602	14	451	17.8	16.4	57.7	8.0
(Simard et al., 2007)	488	55	298	18.3	17.0	56.4	8.3

Table 24: Number of test sentences modified, improved and deteriorated by each submitted run, together with the corresponding percentage of insertions, deletions, substitutions and shifts (case sensitive).

deletions, substitutions, shifts). Looking at these numbers, the following conclusions can be drawn. Although it varies considerably between the submitted runs, the number of modified sentences is quite small. Moreover, a general trend can be observed: the best systems are the most conservative ones. This situation likely reflects the aforementioned data sparsity and coherence issues. A small fraction of the correction patterns found in the training set seems to be applicable also to the test set, and the risk of performing corrections that are either wrong, redundant, or different from those in the reference post-editions is rather high.

From the system point of view, the context in which a learned correction pattern will be applied is crucial. For instance, the same word substitution (*e.g.* "*house*" → "*home*") is not applicable in all contexts. While sometimes it will be necessary (Example 1: "*The house team won the match*"), in some contexts it is optional (Example 2: "*I was in my house*") or wrong (Example 3: "*He worked for a brokerage house*"). Unfortunately, the unnecessary word replacement in Example 2 (human post-editors would likely leave it untouched) would increase the TER of the sentence exactly as in the clearly wrong replacement in Example 3.

From the evaluation point of view, not penalising such correct but unnecessary corrections is also crucial. Similar to MT, where a source sentence can have many valid translations, in the APE task a target sentence can have many valid post-editions. Indeed, nothing prevents that in our evaluation some correct post-editions are considered as "deteriorated" sentences simply because they differ from the human post-editions used as references. As in MT, this well known variability problem might penalise good systems, thus calling for alternative evaluation criteria (*e.g.* based on multiple references or sensitive to paraphrase matches). Interestingly, for all the systems the number of modified sentences is higher than the sum of the improved and the deteriorated ones. Such difference is represented by modified sentences for which the corrections do not yield TER variations. This grey area makes the evaluation problem related to variability even more evident.

The analysis of the edit operations performed by each system is not particularly informative. Similar to the overall performance results, also the proportion of correction types they perform reflects the adoption of the same underlying statistical approach. The distribution of the four types of edit operations is almost identical, with a predominance lexical substitutions (55.7%-57.7%) and rather few phrasal shifts (8.0%-8.6%).

5.5 Lessons learned and outlook

The objectives of this pilot APE task were to: *i)* define a sound evaluation framework for future rounds, *ii)* identify and understand the most critical aspects in terms of data acquisition and system evaluation, *iii)* make an inventory of current approaches, evaluate the state of the art and *iv)* provide a milestone for future studies on the problem. With respect to the first point, improving the evaluation is possible, but no major issues emerged or requested radical changes in future evaluation rounds. For instance, using multiple references or a metric sensitive to paraphrase matches to cope with variability in the post-editing would certainly help.

Concerning the most critical aspects of the evaluation, our analysis highlighted the strong dependence of system results on data repetitiveness/representativeness. This calls into question the actual usability of text material coming

from general domains like `news` and, probably, of post-editions collected from crowdsourced workers (this aspect, however, should be further investigated to draw definite conclusions). Nevertheless, it's worth noting that collecting a large, unpublished, public, domain-specific and professional-quality dataset is a hardly achievable goal that will always require compromise solutions.

Regarding the approaches proposed, this first experience was a conservative but, at the same time, promising first step. Although participants performed the task sharing the same statistical approach to APE, the slight variants they explored allowed them to outperform the widely used monolingual translation technique. Moreover, results' analysis also suggests a possible limitation of this state-of-the-art approach: by always performing all the applicable correction patterns, it runs the risk of deteriorating the input translations that it was supposed to improve. This limitation, common to all the participating systems, is a clue of a major difference between the APE task and the MT framework. In MT the system must always process the entire source sentence by translating all of its words into the target language. In the APE scenario, instead, the system has another option for each word: keeping it untouched. A reasonable (and this year unbeaten) baseline is in fact a system that applies this conservative strategy for all the words. By raising this and other issues as promising research directions, attracting researchers' attention to a challenging application-oriented task, and establishing a sound evaluation framework to measure future advancements, this pilot has substantially achieved its goals, paving the way for future rounds of the APE evaluation exercise.

Acknowledgments

This work was supported in parts by the MosesCore, QT21, EXPERT and CRACKER projects funded by the European Commission (7th Framework Programme and H2020).

We would also like to thank Unbabel for providing the data used in the QE Tasks 1 and 2, and in the APE task.

References

Avramidis, E., Popović, M., and Burchardt, A. (2015). DFKI's experimental hybrid MT system for WMT 2015. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 66–73, Lisboa, Portugal. Association for Computational Linguistics.

Banerjee, S. and Lavie, A. (2005). METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments. In *Proceedings of the ACL 2005 Workshop on Intrinsic and Extrinsic Evaluation Measures for MT and/or Summarization*.

Béchara, H., Ma, Y., and van Genabith, J. (2011). Statistical Post-Editing for a Statistical MT System. In *Proceedings of the 13th Machine Translation Summit*, pages 308–315, Xiamen, China.

Beck, D., Shah, K., Cohn, T., and Specia, L. (2013a). SHEF-Lite: When less is more for translation quality estimation. In *Proceedings of the Eighth Workshop on Statistical Machine Translation*, pages 335–340, Sofia, Bulgaria. Association for Computational Linguistics.

Beck, D., Specia, L., and Cohn, T. (2013b). Reducing annotation effort for quality estimation via active learning. In *51st Annual Meeting of the Association for Computational Linguistics: Short Papers*, ACL, pages 543–548, Sofia, Bulgaria.

Biçici, E. (2013). Referential translation machines for quality estimation. In *Proceedings of the Eighth Workshop on Statistical Machine Translation*, Sofia, Bulgaria.

Biçici, E. and Way, A. (2014). Referential Translation Machines for Predicting Translation Quality. In *Ninth Workshop on Statistical Machine Translation*, pages 313–321, Baltimore, Maryland, USA.

Bicici, E., Liu, Q., and Way, A. (2015). Referential Translation Machines for Predicting Translation Quality and Related Statistics. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 304–308, Lisboa, Portugal. Association for Computational Linguistics.

Blain, F., Schwenk, H., and Senellart, J. (2012). Incremental adaptation using translation information and post-editing analysis. In *International Workshop on Spoken Language Translation (IWSLT)*, pages 234–241, Hong-Kong (China).

Bojar, O., Buck, C., Callison-Burch, C., Federmann, C., Haddow, B., Koehn, P., Monz, C., Post, M., Soricut, R., and Specia, L. (2013).

Findings of the 2013 Workshop on Statistical Machine Translation. In *Proceedings of the Eighth Workshop on Statistical Machine Translation*, pages 1–42, Sofia, Bulgaria. Association for Computational Linguistics.

Bojar, O., Buck, C., Federmann, C., Haddow, B., Koehn, P., Leveling, J., Monz, C., Pecina, P., Post, M., Saint-Amand, H., Soricut, R., Specia, L., and Tamchyna, A. (2014). Findings of the 2014 workshop on statistical machine translation. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 12–58, Baltimore, Maryland, USA. Association for Computational Linguistics.

Bojar, O. and Tamchyna, A. (2015). CUNI in WMT15: Chimera Strikes Again. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 79–83, Lisboa, Portugal. Association for Computational Linguistics.

Callison-Burch, C., Fordyce, C., Koehn, P., Monz, C., and Schroeder, J. (2007). (Meta-) evaluation of machine translation. In *Proceedings of the Second Workshop on Statistical Machine Translation (WMT07)*, Prague, Czech Republic.

Callison-Burch, C., Fordyce, C., Koehn, P., Monz, C., and Schroeder, J. (2008). Further meta-evaluation of machine translation. In *Proceedings of the Third Workshop on Statistical Machine Translation (WMT08)*, Colmbus, Ohio.

Callison-Burch, C., Koehn, P., Monz, C., Peterson, K., Przybocki, M., and Zaidan, O. F. (2010). Findings of the 2010 joint workshop on statistical machine translation and metrics for machine translation. In *Proceedings of the Fourth Workshop on Statistical Machine Translation (WMT10)*, Uppsala, Sweden.

Callison-Burch, C., Koehn, P., Monz, C., Post, M., Soricut, R., and Specia, L. (2012). Findings of the 2012 workshop on statistical machine translation. In *Proceedings of the Seventh Workshop on Statistical Machine Translation*, pages 10–51, Montréal, Canada. Association for Computational Linguistics.

Callison-Burch, C., Koehn, P., Monz, C., and Schroeder, J. (2009). Findings of the 2009 workshop on statistical machine translation. In *Proceedings of the Fourth Workshop on Statistical Machine Translation (WMT09)*, Athens, Greece.

Callison-Burch, C., Koehn, P., Monz, C., and Zaidan, O. (2011). Findings of the 2011 workshop on statistical machine translation. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 22–64, Edinburgh, Scotland.

Camargo de Souza, J. G., González-Rubio, J., Buck, C., Turchi, M., and Negri, M. (2014). Fbk-upv-uedin participation in the wmt14 quality estimation shared-task. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 322–328, Baltimore, Maryland, USA. Association for Computational Linguistics.

Cap, F., Weller, M., Ramm, A., and Fraser, A. (2015). CimS - The CIS and IMS Joint Submission to WMT 2015 addressing morphological and syntactic differences in English to German SMT. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 84–91, Lisboa, Portugal. Association for Computational Linguistics.

Cettolo, M., Bertoldi, N., and Federico, M. (2014). The Repetition Rate of Text as a Predictor of the Effectiveness of Machine Translation Adaptation. In *Proceedings of the 11th Biennial Conference of the Association for Machine Translation in the Americas (AMTA 2014)*, pages 166–179, Vancouver, BC, Canada.

Chatterjee, R., Turchi, M., and Negri, M. (2015a). The FBK Participation in the WMT15 Automatic Post-editing Shared Task. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 210–215, Lisboa, Portugal. Association for Computational Linguistics.

Chatterjee, R., Weller, M., Negri, M., and Turchi, M. (2015b). Exploring the Planet of the APEs: a Comparative Study of State-of-the-art Methods for MT Automatic Post-Editing. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics)*, Beijing, China.

Cho, E., Ha, T.-L., Niehues, J., Herrmann, T., Mediani, M., Zhang, Y., and Waibel, A. (2015). The Karlsruhe Institute of Technology Translation Systems for the WMT 2015. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 92–97, Lisboa, Portugal. Association for Computational Linguistics.

Cohen, J. (1960). A coefficient of agreement for

nominal scales. *Educational and Psychological Measurment*, 20(1):37–46.

Dušek, O., Gomes, L., Novák, M., Popel, M., and Rosa, R. (2015). New Language Pairs in TectoMT. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 98–104, Lisboa, Portugal. Association for Computational Linguistics.

Dyer, C., Lopez, A., Ganitkevitch, J., Weese, J., Ture, F., Blunsom, P., Setiawan, H., Eidelman, V., and Resnik, P. (2010). cdec: A decoder, alignment, and learning framework for finite-state and context-free translation models. In *Proceedings of the Association for Computational Linguistics (ACL)*.

Esplà-Gomis, M., Sánchez-Martínez, F., and Forcada, M. (2015a). UAlacant word-level machine translation quality estimation system at WMT 2015. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 309–315, Lisboa, Portugal. Association for Computational Linguistics.

Esplà-Gomis, M., Sánchez-Martnez, F., and Forcada, M. L. (2015b). Using on-line available sources of bilingual information for word-level machine translation quality estimation. In *18th Annual Conference of the European Association for Machine Translation*, page 1926, Antalya, Turkey.

Federmann, C. (2012). Appraise: An Open-Source Toolkit for Manual Evaluation of Machine Translation Output. *The Prague Bulletin of Mathematical Linguistics (PBML)*, 98:25–35.

Gao, Q. and Vogel, S. (2008). Parallel Implementations of Word Alignment Tool. In *Proceedings of the ACL 2008 Software Engineering, Testing, and Quality Assurance Workshop*, pages 49–57, Columbus, Ohio.

Giménez, J. and Màrquez, L. (2010). Asiya: An Open Toolkit for Automatic Machine Translation (Meta-)Evaluation. *The Prague Bulletin of Mathematical Linguistics*, 94:77–86.

Graham, Y. (2015). Improving Evaluation of Machine Translation Quality Estimation. In *53rd Annual Meeting of the Association for Computational Linguistics and Seventh International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing*, pages 1804–1813, Beijing, China.

Grönroos, S.-A., Virpioja, S., and Kurimo, M. (2015). Tuning Phrase-Based Segmented Translation for a Morphologically Complex Target Language. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 105–111, Lisboa, Portugal. Association for Computational Linguistics.

Gwinnup, J., Anderson, T., Erdmann, G., Young, K., May, C., Kazi, M., Salesky, E., and Thompson, B. (2015). The AFRL-MITLL WMT15 System: Theres More than One Way to Decode It! In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 112–119, Lisboa, Portugal. Association for Computational Linguistics.

Ha, T.-L., Do, Q.-K., Cho, E., Niehues, J., Allauzen, A., Yvon, F., and Waibel, A. (2015). The KIT-LIMSI Translation System for WMT 2015. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 120–125, Lisboa, Portugal. Association for Computational Linguistics.

Haddow, B., Huck, M., Birch, A., Bogoychev, N., and Koehn, P. (2015). The Edinburgh/JHU Phrase-based Machine Translation Systems for WMT 2015. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 126–133, Lisboa, Portugal. Association for Computational Linguistics.

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

Jean, S., Firat, O., Cho, K., Memisevic, R., and Bengio, Y. (2015). Montreal Neural Machine Translation Systems for WMT15. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 134–140, Lisboa, Portugal. Association for Computational Linguistics.

Junczys-Dowmunt, M. and Szal, A. (2011). SyMGiza++: Symmetrized Word Alignment Models for Statistical Machine Translation. In *SIIS*, volume 7053 of *Lecture Notes in Computer Science*, pages 379–390. Springer.

Koehn, P. (2004). Statistical Significance Tests for Machine Translation Evaluation. In Lin, D. and Wu, D., editors, *Proceedings of EMNLP 2004*, pages 388–395, Barcelona, Spain.

Koehn, P. (2005). Europarl: A Parallel Corpus for Statistical Machine Translation. In *MT Summit X*.

Koehn, P. (2010). *Statistical Machine Translation*. Cambridge University Press, New York, NY, USA, 1st edition.

Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., Dyer, C., Bojar, O., Constantin, A., and Herbst, E. (2007). Moses: Open source toolkit for statistical machine translation. In *ACL 2007 Demonstrations*, Prague, Czech Republic.

Koehn, P. and Monz, C. (2006). Manual and automatic evaluation of machine translation between European languages. In *Proceedings of NAACL 2006 Workshop on Statistical Machine Translation*, New York, New York.

Kolachina, P. and Ranta, A. (2015). GF Wide-coverage English-Finnish MT system for WMT 2015. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 141–144, Lisboa, Portugal. Association for Computational Linguistics.

Kreutzer, J., Schamoni, S., and Riezler, S. (2015). QUality Estimation from ScraTCH (QUETCH): Deep Learning for Word-level Translation Quality Estimation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 316–322, Lisboa, Portugal. Association for Computational Linguistics.

Landis, J. R. and Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, 33:159–174.

Langlois, D. (2015). LORIA System for the WMT15 Quality Estimation Shared Task. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 323–329, Lisboa, Portugal. Association for Computational Linguistics.

Lavie, A. and Agarwal, A. (2007). Meteor: An Automatic Metric for MT Evaluation with High Levels of Correlation with Human Judgments. In *Proceedings of the Second Workshop on Statistical Machine Translation*, StatMT '07, pages 228–231.

Liang, P., Taskar, B., and Klein, D. (2006). Alignment by Agreement. In *HLTNAACL*, New York.

Logacheva, V., Hokamp, C., and Specia, L. (2015). Data enhancement and selection strategies for the word-level Quality Estimation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 330–335, Lisboa, Portugal. Association for Computational Linguistics.

Luong, N. Q., Besacier, L., and Lecouteux, B. (2014). LIG System for Word Level QE task at WMT14. In *Ninth Workshop on Statistical Machine Translation*, pages 335–341, Baltimore, Maryland, USA. Association for Computational Linguistics.

Luong, N. Q., Lecouteux, B., and Besacier, L. (2013). LIG system for WMT13 QE task: Investigating the usefulness of features in word confidence estimation for MT. In *Proceedings of the Eighth Workshop on Statistical Machine Translation*, pages 384–389, Sofia, Bulgaria. Association for Computational Linguistics.

Marie, B., Allauzen, A., Burlot, F., Do, Q.-K., Ive, J., knyazeva, e., Labeau, M., Lavergne, T., Löser, K., Pécheux, N., and Yvon, F. (2015). LIMSI@WMT'15 : Translation Task. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 145–151, Lisboa, Portugal. Association for Computational Linguistics.

Och, F. J. (2003). Minimum Error Rate Training in Statistical Machine Translation. In *ACL03*, pages 160–167, Sapporo, Japan.

Padr, L. and Stanilovsky, E. (2012). Freeling 3.0: Towards wider multilinguality. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey.

Pal, S., Naskar, S., and Bandyopadhyay, S. (2013). A Hybrid Word Alignment Model for Phrase-Based Statistical Machine Translation. In *Proceedings of the Second Workshop on Hybrid Approaches to Translation*, pages 94–101, Sofia, Bulgaria.

Pal, S., Naskar, S., and van Genabith, J. (2015a). UdS-Sant: English–German Hybrid Machine Translation System. In *Proceedings of the Tenth*

Workshop on Statistical Machine Translation, pages 152–157, Lisboa, Portugal. Association for Computational Linguistics.

Pal, S., Vela, M., Naskar, S. K., and van Genabith, J. (2015b). USAAR-SAPE: An English–Spanish Statistical Automatic Post-Editing System. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 216–221, Lisboa, Portugal. Association for Computational Linguistics.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, pages 311–318, Morristown, NJ, USA.

Parton, K., Habash, N., McKeown, K., Iglesias, G., and de Gispert, A. (2012). Can Automatic Post-Editing Make MT More Meaningful? In *Proceedings of the 16th Conference of the European Association for Machine Translation (EAMT)*, pages 111–118, Trento, Italy.

Peter, J.-T., Toutounchi, F., Wuebker, J., and Ney, H. (2015). The RWTH Aachen German-English Machine Translation System for WMT 2015. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 158–163, Lisboa, Portugal. Association for Computational Linguistics.

Potet, M., Esperana-Rodier, E., Besacier, L., and Blanchon, H. (2012). Collection of a large database of french-english smt output corrections. In *LREC*, pages 4043–4048. European Language Resources Association (ELRA).

Quernheim, D. (2015). Exact Decoding with Multi Bottom-Up Tree Transducers. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 164–171, Lisboa, Portugal. Association for Computational Linguistics.

Raybaud, S., Langlois, D., and Smali, K. (2011). this sentence is wrong. detecting errors in machine-translated sentences. *Machine Translation*, 25(1):1–34.

Rubino, R., Pirinen, T., Esplà-Gomis, M., Ljubešić, N., Ortiz Rojas, S., Papavassiliou, V., Prokopidis, P., and Toral, A. (2015). Abu-MaTran at WMT 2015 Translation Task: Morphological Segmentation and Web Crawling. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 184–191, Lisboa, Portugal. Association for Computational Linguistics.

Scarton, C., Tan, L., and Specia, L. (2015a). USHEF and USAAR-USHEF participation in the WMT15 QE shared task. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 336–341, Lisboa, Portugal. Association for Computational Linguistics.

Scarton, C., Zampieri, M., Vela, M., van Genabith, J., and Specia, L. (2015b). Searching for Context: a Study on Document-Level Labels for Translation Quality Estimation. In *The 18th Annual Conference of the European Association for Machine Translation*, pages 121–128, Antalya, Turkey.

Schwartz, L., Bryce, B., Geigle, C., Massung, S., Liu, Y., Peng, H., Raja, V., Roy, S., and Upadhyay, S. (2015). The University of Illinois submission to the WMT 2015 Shared Translation Task. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 192–198, Lisboa, Portugal. Association for Computational Linguistics.

Shah, K., Logacheva, V., Paetzold, G., Blain, F., Beck, D., Bougares, F., and Specia, L. (2015). SHEF-NN: Translation Quality Estimation with Neural Networks. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 342–347, Lisboa, Portugal. Association for Computational Linguistics.

Shang, L., Cai, D., and Ji, D. (2015). Strategy-Based Technology for Estimating MT Quality. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 348–352, Lisboa, Portugal. Association for Computational Linguistics.

Simard, M., Goutte, C., and Isabelle, P. (2007). Statistical Phrase-Based Post-Editing. In *Proceedings of the Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL HLT)*, pages 508–515, Rochester, New York.

Snover, M., Dorr, B., Schwartz, R., Micciulla, L., and Makhoul, J. (2006a). A Study of Translation Edit Rate with Targeted Human Annotation. In *Proceedings of Association for Machine Translation in the Americas*, pages 223–231, Cambridge, Massachusetts, USA.

Snover, M., Dorr, B., Schwartz, R., Micciulla, L., and Makhoul, J. (2006b). A study of translation edit rate with targeted human annotation. In *Proceedings of the 7th Biennial Conference of the Association for Machine Translation in the Americas (AMTA-2006)*, Cambridge, Massachusetts.

Specia, L., Paetzold, G., and Scarton, C. (2015). Multi-level Translation Quality Prediction with QuEst++. In *53rd Annual Meeting of the Association for Computational Linguistics and Seventh International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing: System Demonstrations*, pages 115–120, Beijing, China.

Specia, L., Shah, K., de Souza, J. G., and Cohn, T. (2013). QuEst - A translation quality estimation framework. In *51st Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, ACL-2013, pages 79–84, Sofia, Bulgaria.

Stanojević, M., Kamran, A., and Bojar, O. (2015a). Results of the WMT15 Tuning Shared Task. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 274–281, Lisboa, Portugal. Association for Computational Linguistics.

Stanojević, M., Kamran, A., Koehn, P., and Bojar, O. (2015b). Results of the WMT15 Metrics Shared Task. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 256–273, Lisboa, Portugal. Association for Computational Linguistics.

Steele, D., Sim Smith, K., and Specia, L. (2015). Sheffield Systems for the Finnish-English WMT Translation Task. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 172–176, Lisboa, Portugal. Association for Computational Linguistics.

Tezcan, A., Hoste, V., Desmet, B., and Macken, L. (2015). UGENT-LT3 SCATE System for Machine Translation Quality Estimation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 353–360, Lisboa, Portugal. Association for Computational Linguistics.

Tiedemann, J., Ginter, F., and Kanerva, J. (2015). Morphological Segmentation and OPUS for Finnish-English Machine Translation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 177–183, Lisboa, Portugal. Association for Computational Linguistics.

Williams, P., Sennrich, R., Nadejde, M., Huck, M., and Koehn, P. (2015). Edinburgh's Syntax-Based Systems at WMT 2015. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 199–209, Lisboa, Portugal. Association for Computational Linguistics.

Wisniewski, G., Pécheux, N., and Yvon, F. (2015). Why Predicting Post-Edition is so Hard? Failure Analysis of LIMSI Submission to the APE Shared Task. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 222–227, Lisboa, Portugal. Association for Computational Linguistics.

Wisniewski, G., Singh, A. K., Segal, N., and Yvon, F. (2013). Design and Analysis of a Large Corpus of Post-edited Translations: Quality Estimation, Failure Analysis and the Variability of Post-edition. *Machine Translation Summit*, 14:117–124.

	ONLINE-B	UEDIN-JHU	UEDIN-SYNTAX	MONTREAL	ONLINE-A	CU-TECTO	TT-BLEU-MIRA-D	TT-ILLC-UVA	TT-BLEU-MERT	TT-AFRL	TT-USAAR-TUNA	TT-DCU	TT-METEOR-CMU	TT-BLEU-MIRA-SP	TT-HKUST-MEANT	ILLINOIS
ONLINE-B	–	.46†	**.52**	.46†	.39‡	.25‡	.21‡	.21‡	.21‡	.21‡	.20‡	.20‡	.19‡	.17‡	.16‡	.17‡
UEDIN-JHU	**.54**†	–	.48	.47⋆	.44‡	.26‡	.21‡	.22‡	.20‡	.21‡	.20‡	.19‡	.19‡	.19‡	.19‡	.19‡
UEDIN-SYNTAX	.48	**.52**	–	**.51**	.46⋆	.28‡	.21‡	.22‡	.22‡	.21‡	.21‡	.19‡	.18‡	.20‡	.19‡	.17‡
MONTREAL	**.54**†	**.53**⋆	.49	–	.45†	.28‡	.24‡	.25‡	.24‡	.24‡	.25‡	.24‡	.21‡	.20‡	.20‡	.23‡
ONLINE-A	**.61**‡	**.56**‡	**.54**⋆	**.55**†	–	.29‡	.24‡	.26‡	.25‡	.25‡	.24‡	.23‡	.22‡	.23‡	.23‡	.22‡
CU-TECTO	**.75**‡	**.74**‡	**.72**‡	**.72**‡	**.71**‡	–	.48	.47	.47	.46†	.48	.44‡	.43‡	.43‡	.43‡	.41‡
TT-BLEU-MIRA-D	**.79**‡	**.79**‡	**.79**‡	**.76**‡	**.76**‡	**.52**	–	**.51**	.41†	.43⋆	.38†	.43†	.41‡	.39‡	.39‡	.43‡
TT-ILLC-UVA	**.79**‡	**.78**‡	**.78**‡	**.75**‡	**.74**‡	**.53**	.49	–	.48	.47	.45	.41‡	.45⋆	.42‡	.40‡	.42‡
TT-BLEU-MERT	**.79**‡	**.80**‡	**.78**‡	**.76**‡	**.75**‡	**.53**	**.59**†	**.52**	–	**.51**	.48	.44†	.45†	.41‡	.40‡	.41‡
TT-AFRL	**.79**‡	**.79**‡	**.79**‡	**.76**‡	**.75**‡	**.54**†	**.57**⋆	**.53**	.49	–	.49	.45⋆	.43†	.42‡	.42‡	.41‡
TT-USAAR-TUNA	**.80**‡	**.80**‡	**.79**‡	**.75**‡	**.76**‡	**.52**	**.62**†	**.55**	**.52**	**.51**	–	.45⋆	.45†	.41‡	.41‡	.42‡
TT-DCU	**.80**‡	**.81**‡	**.81**‡	**.76**‡	**.77**‡	**.56**‡	**.57**†	**.59**‡	**.56**†	**.55**⋆	**.55**⋆	–	.47	.45†	.44†	.45†
TT-METEOR-CMU	**.81**‡	**.81**‡	**.82**‡	**.79**‡	**.78**‡	**.57**‡	**.59**‡	**.55**⋆	**.55**†	**.57**†	**.55**†	**.53**	–	.48	.49	.48
TT-BLEU-MIRA-SP	**.83**‡	**.81**‡	**.80**‡	**.80**‡	**.77**‡	**.57**‡	**.61**‡	**.58**‡	**.59**‡	**.58**‡	**.59**‡	**.55**†	**.52**	–	**.53**	.50
TT-HKUST-MEANT	**.84**‡	**.81**‡	**.81**‡	**.80**‡	**.77**‡	**.57**‡	**.61**‡	**.60**‡	**.60**‡	**.58**‡	**.59**‡	**.56**†	**.51**	.47	–	.48
ILLINOIS	**.82**‡	**.81**‡	**.83**‡	**.77**‡	**.78**‡	**.59**‡	**.57**‡	**.58**‡	**.59**‡	**.59**‡	**.58**‡	**.55**†	**.52**	.50	**.52**	–
score	.61	.57	.53	.51	.43	-.12	-.18	-.18	-.19	-.21	-.22	-.26	-.29	-.32	-.32	-.35
rank	1	2	3-4	3-4	5	6	7-9	7-10	7-11	8-11	9-11	12-13	13-15	13-15	13-15	15-16

Table 25: Head to head comparison, ignoring ties, for Czech-English systems

A Pairwise System Comparisons by Human Judges

Tables 25–34 show pairwise comparisons between systems for each language pair. The numbers in each of the tables' cells indicate the percentage of times that the system in that column was judged to be better than the system in that row, ignoring ties. Bolding indicates the winner of the two systems.

Because there were so many systems and data conditions the significance of each pairwise comparison needs to be quantified. We applied the Sign Test to measure which comparisons indicate genuine differences (rather than differences that are attributable to chance). In the following tables ⋆ indicates statistical significance at $p \leq 0.10$, † indicates statistical significance at $p \leq 0.05$, and ‡ indicates statistical significance at $p \leq 0.01$, according to the Sign Test.

Each table contains final rows showing how likely a system would win when paired against a randomly selected system (the expected win ratio score) and the rank range according bootstrap resampling ($p \leq 0.05$). Gray lines separate clusters based on non-overlapping rank ranges.

	CU-CHIMERA	ONLINE-B	UEDIN-JHU	MONTREAL	ONLINE-A	UEDIN-SYNTAX	CU-TECTO	COMMERCIAL1	TT-DCU	TT-AFRL	TT-BLEU-MIRA-D	TT-USAAR-TUNA	TT-BLEU-MERT	TT-METEOR-CMU	TT-BLEU-MIRA-SP
CU-CHIMERA	–	.42‡	.43‡	.44‡	.38‡	.33‡	.29‡	.27‡	.15‡	.15‡	.15‡	.14‡	.14‡	.11‡	.10‡
ONLINE-B	**.58**‡	–	.50	.50	.44‡	.40‡	.37‡	.32‡	.16‡	.17‡	.17‡	.17‡	.16‡	.13‡	.08‡
UEDIN-JHU	**.57**‡	.50	–	**.51**	.44‡	.39‡	.41‡	.35‡	.18‡	.18‡	.18‡	.18‡	.16‡	.13‡	.10‡
MONTREAL	**.56**‡	.50	.49	–	.46†	.43‡	.39‡	.36‡	.22‡	.21‡	.21‡	.21‡	.19‡	.19‡	.16‡
ONLINE-A	**.62**‡	**.56**‡	**.56**‡	**.54**†	–	.43‡	.40‡	.36‡	.20‡	.19‡	.20‡	.18‡	.17‡	.15‡	.12‡
UEDIN-SYNTAX	**.67**‡	**.60**‡	**.61**‡	**.57**‡	**.57**‡	–	.48	.43‡	.25‡	.25‡	.26‡	.25‡	.23‡	.23‡	.17‡
CU-TECTO	**.71**‡	**.62**‡	**.59**‡	**.61**‡	**.60**‡	**.52**	–	.44‡	.29‡	.30‡	.28‡	.28‡	.28‡	.23‡	.17‡
COMMERCIAL1	**.73**‡	**.68**‡	**.65**‡	**.64**‡	**.64**‡	**.57**‡	**.56**‡	–	.29‡	.28‡	.28‡	.27‡	.27‡	.22‡	.18‡
TT-DCU	**.85**‡	**.84**‡	**.82**‡	**.78**‡	**.80**‡	**.75**‡	**.71**‡	**.71**‡	–	**.52**	.48	.45†	.40‡	.36‡	.27‡
TT-AFRL	**.85**‡	**.83**‡	**.82**‡	**.79**‡	**.81**‡	**.75**‡	**.70**‡	**.72**‡	.48	–	.49	.46⋆	.37‡	.33‡	.29‡
TT-BLEU-MIRA-D	**.85**‡	**.83**‡	**.82**‡	**.79**‡	**.80**‡	**.74**‡	**.72**‡	**.72**‡	**.52**	**.51**	–	.39‡	.36‡	.36‡	.27‡
TT-USAAR-TUNA	**.86**‡	**.83**‡	**.82**‡	**.79**‡	**.82**‡	**.75**‡	**.72**‡	**.73**‡	**.55**†	**.54**⋆	**.61**‡	–	.36‡	.37‡	.28‡
TT-BLEU-MERT	**.86**‡	**.84**‡	**.84**‡	**.81**‡	**.83**‡	**.77**‡	**.72**‡	**.73**‡	**.60**‡	**.63**‡	**.64**‡	**.64**‡	–	.39‡	.28‡
TT-METEOR-CMU	**.89**‡	**.87**‡	**.87**‡	**.81**‡	**.85**‡	**.77**‡	**.77**‡	**.78**‡	**.64**‡	**.67**‡	**.64**‡	**.63**‡	**.61**‡	–	.32‡
TT-BLEU-MIRA-SP	**.90**‡	**.92**‡	**.90**‡	**.84**‡	**.88**‡	**.83**‡	**.83**‡	**.82**‡	**.73**‡	**.71**‡	**.73**‡	**.72**‡	**.72**‡	**.68**‡	–
score	.68	.51	.50	.46	.42	.26	.20	.11	-.34	-.34	-.34	-.37	-.40	-.56	-.80
rank	1	2-3	2-3	4	5	6	7	8	9-11	9-11	9-11	12	13	14	15

Table 26: Head to head comparison, ignoring ties, for English-Czech systems

	ONLINE-B	UEDIN-JHU	ONLINE-A	UEDIN-SYNTAX	KIT	RWTH	MONTREAL	ILLINOIS	DFKI	ONLINE-C	ONLINE-F	MACAU	ONLINE-E
ONLINE-B	–	.41‡	.43‡	.39‡	.39‡	.33‡	.38‡	.25‡	.26‡	.27‡	.26‡	.19‡	.22‡
UEDIN-JHU	**.59**‡	–	**.51**	.46⋆	.45†	.43†	.44†	.31‡	.33‡	.36‡	.30‡	.28‡	.27‡
ONLINE-A	**.57**‡	.49	–	**.52**	**.53**	.48	.44‡	.36‡	.32‡	.31‡	.28‡	.29‡	.26‡
UEDIN-SYNTAX	**.61**‡	**.54**⋆	.48	–	.49	.48	.45†	.23‡	.33‡	.34‡	.35‡	.27‡	.26‡
KIT	**.61**‡	**.55**†	.47	**.51**	–	.47	.46⋆	.35‡	.38‡	.36‡	.35‡	.26‡	.32‡
RWTH	**.67**‡	**.57**†	**.52**	**.52**	**.53**	–	.46⋆	.38‡	.39‡	.40‡	.36‡	.31‡	.35‡
MONTREAL	**.62**‡	**.56**†	**.56**‡	**.55**†	**.54**⋆	**.54**⋆	–	.42‡	.43‡	.41‡	.35‡	.32‡	.34‡
ILLINOIS	**.75**‡	**.69**‡	**.64**‡	**.77**‡	**.65**‡	**.62**‡	**.58**‡	–	.48	.49	.48	.38‡	.42‡
DFKI	**.74**‡	**.67**‡	**.68**‡	**.67**‡	**.62**‡	**.61**‡	**.57**‡	**.52**	–	.43†	.46⋆	.39‡	.37‡
ONLINE-C	**.73**‡	**.64**‡	**.69**‡	**.66**‡	**.64**‡	**.60**‡	**.59**‡	**.51**	**.57**†	–	.46⋆	.42‡	.39‡
ONLINE-F	**.74**‡	**.70**‡	**.72**‡	**.65**‡	**.65**‡	**.64**‡	**.64**‡	**.52**	**.54**⋆	**.54**⋆	–	.44‡	.40‡
MACAU	**.81**‡	**.72**‡	**.71**‡	**.73**‡	**.74**‡	**.69**‡	**.68**‡	**.62**‡	**.61**‡	**.58**‡	**.56**‡	–	.50
ONLINE-E	**.78**‡	**.73**‡	**.74**‡	**.74**‡	**.68**‡	**.65**‡	**.66**‡	**.58**‡	**.63**‡	**.61**‡	**.60**‡	.50	–
score	.56	.31	.29	.25	.22	.14	.09	-.17	-.17	-.22	-.30	-.48	-.54
rank	1	2-3	2-4	3-5	4-5	6-7	6-7	8-10	8-10	9-10	11	12-13	12-13

Table 27: Head to head comparison, ignoring ties, for German-English systems

	UEDIN-SYNTAX	MONTREAL	PROMT-RULE	ONLINE-A	ONLINE-B	KIT-LIMSI	UEDIN-JHU	ONLINE-F	ONLINE-C	KIT	CIMS	DFKI	ONLINE-E	UDS-SANT	ILLINOIS	IMS
UEDIN-SYNTAX	–	**.52**	.47	.48	.42‡	.36‡	.36‡	.33‡	.37‡	.32‡	.29‡	.32‡	.31‡	.33‡	.19‡	.21‡
MONTREAL	.48	–	.47	.44†	.41‡	.35‡	.35‡	.42‡	.37‡	.35‡	.33‡	.33‡	.37‡	.35‡	.24‡	.27‡
PROMT-RULE	**.53**	**.53**	–	.46⋆	.45†	.46⋆	.40‡	.35‡	.42‡	.41‡	.37‡	.36‡	.33‡	.37‡	.29‡	.24‡
ONLINE-A	**.52**	**.56**†	**.54**⋆	–	.40‡	.43†	.37‡	.42‡	.39‡	.39‡	.41‡	.36‡	.36‡	.33‡	.27‡	.28‡
ONLINE-B	**.58**‡	**.59**‡	**.55**†	**.60**‡	–	.45†	.45†	.45†	.44†	.39‡	.42‡	.37‡	.41‡	.35‡	.29‡	.32‡
KIT-LIMSI	**.64**‡	**.65**‡	**.54**⋆	**.57**†	**.55**†	–	**.52**	.49	.44†	.40‡	.47	.38‡	.39‡	.37‡	.29‡	.30‡
UEDIN-JHU	**.64**‡	**.65**‡	**.60**‡	**.63**‡	**.55**†	.48	–	.47	**.51**	.46⋆	.43†	.45⋆	.44†	.41‡	.34‡	.30‡
ONLINE-F	**.67**‡	**.58**‡	**.65**‡	**.58**‡	**.55**†	**.51**	**.53**	–	.50	.46⋆	.49	.44†	.46⋆	.39‡	.36‡	.36‡
ONLINE-C	**.63**‡	**.63**‡	**.58**‡	**.61**‡	**.56**†	**.56**†	.49	.50	–	**.52**	.48	.45	.40‡	.42‡	.36‡	.35‡
KIT	**.68**‡	**.65**‡	**.59**‡	**.61**‡	**.61**‡	**.60**‡	**.54**⋆	**.54**⋆	.48	–	**.51**	.43‡	.47	.37‡	.35‡	.33‡
CIMS	**.71**‡	**.67**‡	**.62**‡	**.59**‡	**.58**‡	**.53**	**.57**†	**.51**	**.52**	.49	–	.47	.45†	.44†	.23‡	.34‡
DFKI	**.68**‡	**.67**‡	**.64**‡	**.64**‡	**.63**‡	**.62**‡	**.55**⋆	**.56**†	**.55**	**.57**‡	**.53**	–	.50	.44†	.41‡	.36‡
ONLINE-E	**.69**‡	**.63**‡	**.67**‡	**.64**‡	**.59**‡	**.61**‡	**.56**†	**.54**⋆	**.60**‡	**.53**	**.55**†	.50	–	.45†	.42‡	.38‡
UDS-SANT	**.67**‡	**.65**‡	**.63**‡	**.67**‡	**.65**‡	**.63**‡	**.59**‡	**.61**‡	**.58**‡	**.63**‡	**.56**†	**.56**†	**.55**†	–	.45†	.41‡
ILLINOIS	**.81**‡	**.76**‡	**.71**‡	**.73**‡	**.71**‡	**.71**‡	**.66**‡	**.64**‡	**.64**‡	**.65**‡	**.77**‡	**.59**‡	**.58**‡	**.55**†	–	.48
IMS	**.79**‡	**.73**‡	**.76**‡	**.72**‡	**.68**‡	**.70**‡	**.70**‡	**.64**‡	**.65**‡	**.67**‡	**.66**‡	**.64**‡	**.62**‡	**.59**‡	**.52**	–
score	.35	.33	.26	.23	.14	.08	.03	.00	-.00	-.01	-.03	-.13	-.13	-.23	-.40	-.50
rank	1-2	1-2	3-4	3-4	5	6	7-9	7-11	7-11	8-11	9-11	12-13	12-13	14	15	16

Table 28: Head to head comparison, ignoring ties, for English-German systems

	ONLINE-B	LIMSI-CNRS	UEDIN-JHU	MACAU	ONLINE-A	ONLINE-F	ONLINE-E
ONLINE-B	–	.50	.49	.47†	.44‡	.35‡	.22‡
LIMSI-CNRS	.50	–	.49	.46‡	.45‡	.37‡	.25‡
UEDIN-JHU	**.51**	**.51**	–	.47†	.46†	.35‡	.26‡
MACAU	**.53**†	**.54**‡	**.53**†	–	.48	.39‡	.28‡
ONLINE-A	**.56**‡	**.55**‡	**.54**†	**.52**	–	.38‡	.26‡
ONLINE-F	**.65**‡	**.63**‡	**.65**‡	**.61**‡	**.62**‡	–	.37‡
ONLINE-E	**.78**‡	**.75**‡	**.74**‡	**.72**‡	**.74**‡	**.63**‡	–
score	.49	.44	.41	.27	.22	-.42	-1.43
rank	1-2	1-3	1-3	4-5	4-5	6	7

Table 29: Head to head comparison, ignoring ties, for French-English systems

	LIMSI-CNRS	ONLINE-A	UEDIN-JHU	ONLINE-B	CIMS	ONLINE-F	ONLINE-E
LIMSI-CNRS	–	.45‡	.44‡	.45‡	.38‡	.36‡	.28‡
ONLINE-A	**.55**‡	–	.49	.48⋆	.45‡	.37‡	.32‡
UEDIN-JHU	**.56**‡	**.51**	–	.48⋆	.44‡	.41‡	.31‡
ONLINE-B	**.55**‡	**.52**⋆	**.52**⋆	–	.46‡	.40‡	.31‡
CIMS	**.62**‡	**.55**‡	**.56**‡	**.54**‡	–	.45‡	.36‡
ONLINE-F	**.64**‡	**.63**‡	**.59**‡	**.60**‡	**.55**‡	–	.41‡
ONLINE-E	**.72**‡	**.68**‡	**.69**‡	**.69**‡	**.64**‡	**.59**‡	–
score	.54	.30	.25	.21	-.00	-.33	-.97
rank	1	2-3	2-4	3-4	5	6	7

Table 30: Head to head comparison, ignoring ties, for English-French systems

	ONLINE-B	PROMT-SMT	ONLINE-A	UU-UNC	UEDIN-JHU	ABUMATRAN-COMB	UEDIN-SYNTAX	ILLINOIS	ABUMATRAN-HFS	MONTREAL	ABUMATRAN	LIMSI	SHEFFIELD	SHEFF-STEM
ONLINE-B	–	.36‡	.32‡	.35‡	.29‡	.35‡	.35‡	.29‡	.29‡	.31‡	.17‡	.18‡	.15‡	.15‡
PROMT-SMT	**.64**‡	–	.49	.49	.48	.46	.44†	.43†	.36‡	.34‡	.25‡	.28‡	.25‡	.24‡
ONLINE-A	**.68**‡	**.51**	–	.50	.46	.42‡	.47	.45⋆	.38‡	.40‡	.32‡	.30‡	.25‡	.25‡
UU-UNC	**.65**‡	**.51**	.50	–	.50	.45⋆	.47	.47	.37‡	.34‡	.35‡	.26‡	.26‡	.26‡
UEDIN-JHU	**.71**‡	**.52**	**.54**	.50	–	.49	.50	.47	.42‡	.38‡	.33‡	.31‡	.24‡	.24‡
ABUMATRAN-COMB	**.65**‡	**.54**	**.58**‡	**.55**⋆	**.51**	–	.49	.46	.33‡	.38‡	.23‡	.33‡	.24‡	.24‡
UEDIN-SYNTAX	**.65**‡	**.56**†	**.53**	**.53**	.50	**.51**	–	.44†	.41‡	.42‡	.36‡	.29‡	.30‡	.30‡
ILLINOIS	**.71**‡	**.57**†	**.55**⋆	**.53**	**.53**	**.54**	**.56**†	–	.45⋆	.41‡	.37‡	.33‡	.28‡	.27‡
ABUMATRAN-HFS	**.71**‡	**.64**‡	**.62**‡	**.63**‡	**.58**‡	**.67**‡	**.59**‡	**.55**⋆	–	.42‡	.43†	.38‡	.38‡	.37‡
MONTREAL	**.69**‡	**.66**‡	**.60**‡	**.66**‡	**.62**‡	**.62**‡	**.58**‡	**.59**‡	**.58**‡	–	.48	.43†	.39‡	.39‡
ABUMATRAN	**.83**‡	**.75**‡	**.68**‡	**.65**‡	**.67**‡	**.77**‡	**.64**‡	**.63**‡	**.57**†	**.52**	–	.46	.41‡	.41‡
LIMSI	**.82**‡	**.72**‡	**.70**‡	**.74**‡	**.69**‡	**.67**‡	**.71**‡	**.67**‡	**.62**‡	**.57**†	**.54**	–	**.52**	**.52**
SHEFFIELD	**.85**‡	**.75**‡	**.75**‡	**.74**‡	**.76**‡	**.76**‡	**.70**‡	**.72**‡	**.62**‡	**.61**‡	**.59**‡	.48	–	.00
SHEFF-STEM	**.85**‡	**.76**‡	**.75**‡	**.74**‡	**.76**‡	**.76**‡	**.70**‡	**.73**‡	**.63**‡	**.61**‡	**.59**‡	.48	**1.00**	–
score	.67	.28	.24	.23	.18	.16	.14	.08	-.08	-.17	-.27	-.43	-.51	-.52
rank	1	2-4	2-5	2-5	4-7	5-7	5-8	7-8	9	10	11	12-13	13-14	13-14

Table 31: Head to head comparison, ignoring ties, for Finnish-English systems

.

	ONLINE-B	ONLINE-A	UU-UNC	ABUMATRAN-UNC-COM	ABUMATRAN-COMB	AALTO	UEDIN-SYNTAX	ABUMATRAN-UNC	CMU	CHALMERS
ONLINE-B	–	.40‡	.31‡	.28‡	.24‡	.26‡	.25‡	.25‡	.23‡	.18‡
ONLINE-A	**.60**‡	–	.40‡	.41‡	.36‡	.33‡	.36‡	.34‡	.29‡	.26‡
UU-UNC	**.69**‡	**.60**‡	–	.47⋆	.43‡	.41‡	.37‡	.41‡	.36‡	.27‡
ABUMATRAN-UNC-COM	**.72**‡	**.59**‡	**.53**⋆	–	.45†	.46†	.45‡	.40‡	.41‡	.32‡
ABUMATRAN-COMB	**.76**‡	**.64**‡	**.57**‡	**.55**†	–	.45†	.46†	.47	.42‡	.34‡
AALTO	**.74**‡	**.67**‡	**.59**‡	**.54**†	**.55**†	–	.47	.47⋆	.46†	.33‡
UEDIN-SYNTAX	**.75**‡	**.64**‡	**.63**‡	**.55**‡	**.54**†	**.53**	–	.49	.44‡	.34‡
ABUMATRAN-UNC	**.75**‡	**.66**‡	**.59**‡	**.60**‡	**.53**	**.53**⋆	**.51**	–	.50	.39‡
CMU	**.77**‡	**.71**‡	**.64**‡	**.59**‡	**.58**‡	**.54**†	**.56**‡	.50	–	.40‡
CHALMERS	**.82**‡	**.74**‡	**.73**‡	**.68**‡	**.66**‡	**.67**‡	**.66**‡	**.61**‡	**.60**‡	–
score	1.06	.54	.21	.04	-.05	-.14	-.18	-.21	-.34	-.92
rank	1	2	3	4	5	6-7	6-8	6-8	9	10

Table 32: Head to head comparison, ignoring ties, for English-Finnish systems

	ONLINE-G	ONLINE-B	PROMT-RULE	AFRL-MIT-PB	AFRL-MIT-FAC	ONLINE-A	AFRL-MIT-H	LIMSI-NCODE	UEDIN-SYNTAX	UEDIN-JHU	USAAR-GACHA	USAAR-GACHA	ONLINE-F
ONLINE-G	–	.40‡	.39‡	.35‡	.38‡	.38‡	.34‡	.32‡	.36‡	.33‡	.25‡	.24‡	.21‡
ONLINE-B	**.60**‡	–	.41‡	.44†	.42‡	.43†	.40‡	.38‡	.37‡	.35‡	.29‡	.31‡	.22‡
PROMT-RULE	**.61**‡	**.59**‡	–	.46⋆	.47	**.51**	.47	.47	.46†	.48	.40‡	.41‡	.24‡
AFRL-MIT-PB	**.65**‡	**.56**†	**.54**⋆	–	.49	**.53**	.46	.48	.44†	.44†	.33‡	.33‡	.29‡
AFRL-MIT-FAC	**.62**‡	**.58**‡	**.53**	**.51**	–	.50	.48	.45†	.45†	.46⋆	.34‡	.28‡	.29‡
ONLINE-A	**.62**‡	**.57**†	.49	.47	.50	–	.44†	.49	.48	.44†	.36‡	.36‡	.29‡
AFRL-MIT-H	**.66**‡	**.60**‡	**.53**	**.54**	**.52**	**.56**†	–	.50	.47	.46⋆	.40‡	.34‡	.30‡
LIMSI-NCODE	**.68**‡	**.62**‡	**.53**	**.52**	**.55**†	**.51**	.50	–	.48	.49	.43†	.39‡	.33‡
UEDIN-SYNTAX	**.64**‡	**.63**‡	**.54**†	**.56**†	**.55**†	**.52**	**.53**	**.52**	–	.48	.40‡	.40‡	.34‡
UEDIN-JHU	**.67**‡	**.65**‡	**.52**	**.56**†	**.54**⋆	**.56**†	**.54**⋆	**.51**	**.52**	–	.36‡	.38‡	.33‡
USAAR-GACHA	**.75**‡	**.71**‡	**.60**‡	**.67**‡	**.66**‡	**.64**‡	**.60**‡	**.57**†	**.60**‡	**.64**‡	–	.44⋆	.38‡
USAAR-GACHA	**.76**‡	**.69**‡	**.59**‡	**.67**‡	**.72**‡	**.64**‡	**.66**‡	**.61**‡	**.60**‡	**.62**‡	**.56**⋆	–	.40‡
ONLINE-F	**.79**‡	**.78**‡	**.76**‡	**.71**‡	**.71**‡	**.71**‡	**.70**‡	**.67**‡	**.66**‡	**.67**‡	**.62**‡	**.60**‡	–
score	.49	.31	.12	.11	.11	.10	.05	.01	-.02	-.03	-.21	-.27	-.78
rank	1	2	3-6	3-6	3-6	3-7	6-8	7-10	8-10	8-10	11	12	13

Table 33: Head to head comparison, ignoring ties, for Russian-English systems

	PROMT-RULE	ONLINE-G	ONLINE-B	LIMSI-NCODE	ONLINE-A	UEDIN-JHU	UEDIN-SYNTAX	USAAR-GACHA	USAAR-GACHA	ONLINE-F
PROMT-RULE	–	.39‡	.29‡	.27‡	.28‡	.26‡	.21‡	.21‡	.21‡	.07‡
ONLINE-G	**.61**‡	–	.40‡	.38‡	.33‡	.36‡	.30‡	.25‡	.24‡	.12‡
ONLINE-B	**.71**‡	**.60**‡	–	.49	.44‡	.44‡	.37‡	.33‡	.32‡	.19‡
LIMSI-NCODE	**.73**‡	**.62**‡	**.51**	–	.49	.46†	.38‡	.36‡	.34‡	.22‡
ONLINE-A	**.72**‡	**.67**‡	**.56**‡	**.51**	–	.47⋆	.43‡	.40‡	.36‡	.18‡
UEDIN-JHU	**.74**‡	**.64**‡	**.56**‡	**.54**†	**.53**⋆	–	.46†	.40‡	.36‡	.25‡
UEDIN-SYNTAX	**.79**‡	**.70**‡	**.63**‡	**.62**‡	**.57**‡	**.54**†	–	.45†	.39‡	.25‡
USAAR-GACHA	**.79**‡	**.75**‡	**.67**‡	**.64**‡	**.60**‡	**.60**‡	**.55**†	–	.46	.29‡
USAAR-GACHA	**.79**‡	**.76**‡	**.68**‡	**.66**‡	**.64**‡	**.64**‡	**.61**‡	**.54**	–	.28‡
ONLINE-F	**.93**‡	**.88**‡	**.81**‡	**.78**‡	**.82**‡	**.75**‡	**.75**‡	**.71**‡	**.72**‡	–
score	1.01	.52	.21	.12	.07	.01	-.13	-.27	-.33	-1.21
rank	1	2	3	4-5	4-5	6	7	8	9	10

Table 34: Head to head comparison, ignoring ties, for English-Russian systems