Findings of the 2013 Workshop on Statistical Machine Translation

Ondřej Bojar
Charles University in Prague

Christian Buck
University of Edinburgh

Chris Callison-Burch
University of Pennsylvania

Christian Federmann
Saarland University

Barry Haddow
University of Edinburgh

Philipp Koehn
University of Edinburgh

Christof Monz
University of Amsterdam

Matt Post
Johns Hopkins University

Radu Soricut
Google

Lucia Specia
University of Sheffield

Abstract

We present the results of the WMT13 shared tasks, which included a translation task, a task for run-time estimation of machine translation quality, and an unofficial metrics task. This year, 143 machine translation systems were submitted to the ten translation tasks from 23 institutions. An additional 6 anonymized systems were included, and were then evaluated both automatically and manually, in our largest manual evaluation to date. The quality estimation task had four subtasks, with a total of 14 teams, submitting 55 entries.

1 Introduction

We present the results of the shared tasks of the Workshop on Statistical Machine Translation (WMT) held at ACL 2013. This workshop builds on seven previous WMT workshops (Koehn and Monz, 2006; Callison-Burch et al., 2007, 2008, 2009, 2010, 2011, 2012).

This year we conducted three official tasks: a translation task, a human evaluation of translation results, and a quality estimation task.[1] 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, Spanish, and Russian. The Russian translation tasks were new this year, and were also the most popular. The system outputs for each task were evaluated both automatically and manually.

The human evaluation task (§3) involves asking human judges to rank sentences output by anonymized systems. We obtained large numbers of rankings from two groups: researchers (who contributed evaluations proportional to the number of tasks they entered) and workers on Amazon's Mechanical Turk (who were paid). This year's effort was our largest yet by a wide margin; we managed to collect an order of magnitude more judgments than in the past, allowing us to achieve statistical significance on the majority of the pairwise system rankings. This year, we are also clustering the systems according to these significance results, instead of presenting a total ordering over systems.

The focus of the quality estimation task (§6) is to produce real-time estimates of sentence- or word-level machine translation quality. This task has potential usefulness in a range of settings, such as prioritizing output for human post-editing, or selecting the best translations from a number of systems. This year the following subtasks were proposed: prediction of percentage of word edits necessary to fix a sentence, ranking of up to five alternative translations for a given source sentence, prediction of post-editing time for a sentence, and prediction of word-level scores for a given translation (correct/incorrect and types of edits). The datasets included English-Spanish and German-English news translations produced by a number of machine translation systems. This marks the second year we have conducted this task.

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 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 machine translation, system combination, and automatic evaluation or prediction of translation quality.

[1] The traditional metrics task is evaluated in a separate paper (Macháček and Bojar, 2013).

[2] `http://statmt.org/wmt13/results.html`

Proceedings of the Eighth Workshop on Statistical Machine Translation, pages 1–44,
Sofia, Bulgaria, August 8-9, 2013 ©2013 Association for Computational Linguistics

2 Overview of the Translation Task

The recurring task of the workshop examines translation between English and five other languages: German, Spanish, French, Czech, and — new this year — Russian. We created a test set for each language pair by translating newspaper articles and provided training data.

2.1 Test data

The test data for this year's task was selected from news stories from online sources. A total of 52 articles were selected, in roughly equal amounts from a variety of Czech, English, French, German, Spanish, and Russian news sites:[3]

Czech: aktuálně.cz (1), CTK (1), deník (1), iDNES.cz (3), lidovky.cz (1), Novinky.cz (2)

French: Cyber Presse (3), Le Devoir (1), Le Monde (3), Liberation (2)

Spanish: ABC.es (2), BBC Spanish (1), El Periodico (1), Milenio (3), Noroeste (1), Primera Hora (3)

English: BBC (2), CNN (2), Economist (1), Guardian (1), New York Times (2), The Telegraph (1)

German: Der Standard (1), Deutsche Welle (1), FAZ (1), Frankfurter Rundschau (2), Welt (2)

Russian: AIF (2), BBC Russian (2), Izvestiya (1), Rosbalt (1), Vesti (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.[4] 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[5], United Nations, French-English 10^9 corpus, CzEng), some were updated (News Commentary, monolingual data), and new corpora were added (Common Crawl (Smith et al., 2013), Russian-English parallel data provided by Yandex, Russian-English Wikipedia Headlines provided by CMU).

Some statistics about the training materials are given in Figure 1.

2.3 Submitted systems

We received 143 submissions from 23 institutions. The participating institutions and their entry names are listed in Table 1; each system did not necessarily appear in all translation tasks. We also included three commercial off-the-shelf MT systems and three online statistical MT systems,[6] 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

As with past workshops, we contend that automatic measures of machine translation quality are an imperfect substitute for human assessments. We therefore conduct a manual evaluation of the system outputs and define its results to be the principal ranking of the workshop. In this section, we describe how we collected this data and compute the results, and then present the official results of the ranking.

We run the evaluation campaign using an updated version of Appraise (Federmann, 2012); the tool has been extended to support collecting judgments using Amazon's Mechanical Turk, replacing the annotation system used in previous WMTs. The software, including all changes made for this year's workshop, is available from GitHub.[7]

This year differs from prior years in a few important ways:

- We collected about ten times more judgments that we have in the past, using judgments from both participants in the shared task and non-experts hired on Amazon's Mechanical Turk.
- Instead of presenting a total ordering of systems for each pair, we cluster them and report a ranking over the clusters.

[3] 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.

[4] `http://www.yandex.com/`

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

[6] Thanks to Hervé Saint-Amand and Martin Popel for harvesting these entries.

[7] `https://github.com/cfedermann/Appraise`

Europarl Parallel Corpus

	Spanish ↔ English		French ↔ English		German ↔ English		Czech ↔ English	
Sentences	1,965,734		2,007,723		1,920,209		646,605	
Words	56,895,229	54,420,026	60,125,563	55,642,101	50,486,398	53,008,851	14,946,399	17,376,433
Distinct words	176,258	117,481	140,915	118,404	381,583	115,966	172,461	63,039

News Commentary Parallel Corpus

	Spanish ↔ English		French ↔ English		German ↔ English		Czech ↔ English		Russian ↔ English	
Sentences	174,441		157,168		178,221		140,324		150,217	
Words	5,116,388	4,520,796	4,928,135	4,066,721	4,597,904	4,541,058	3,206,423	3,507,249	3,841,950	4,008,949
Distinct words	84,273	61,693	69,028	58,295	142,461	61,761	138,991	54,270	145,997	57,991

Common Crawl Parallel Corpus

	Spanish ↔ English		French ↔ English		German ↔ English		Czech ↔ English		Russian ↔ English	
Sentences	1,845,286		3,244,152		2,399,123		161,838		878,386	
Words	49,561,060	46,861,758	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	710,755	640,778	889,291	859,017	1,640,835	823,480	210,170	128,212	764,203	432,062

United Nations Parallel Corpus

	Spanish ↔ English		French ↔ English	
Sentences	11,196,913		12,886,831	
Words	318,788,686	365,127,098	411,916,781	360,341,450
Distinct words	593,567	581,339	565,553	666,077

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

Yandex 1M Parallel Corpus

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

Wiki Headlines Parallel Corpus

	Russian ↔ English	
Sentences	514,859	
Words	1,191,474	1,230,644
Distinct words	282,989	251,328

Europarl Language Model Data

	English	Spanish	French	German	Czech
Sentence	2,218,201	2,123,835	2,190,579	2,176,537	668,595
Words	59,848,044	60,476,282	63,439,791	53,534,167	14,946,399
Distinct words	123,059	181,837	145,496	394,781	172,461

News Language Model Data

	English	Spanish	French	German	Czech	Russian
Sentence	68,521,621	13,384,314	21,195,476	54,619,789	27,540,749	19,912,911
Words	1,613,778,461	386,014,234	524,541,570	983,818,841	456,271,247	351,595,790
Distinct words	3,392,137	1,163,825	1,590,187	6,814,953	2,655,813	2,195,112

News Test Set

	English	Spanish	French	German	Czech	Russian
Sentences	3000					
Words	64,810	73,659	73,659	63,412	57,050	58,327
Distinct words	8,935	10,601	11,441	12,189	15,324	15,736

Figure 1: 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
BALAGUR	Yandex School of Data Analysis (Borisov et al., 2013)
CMU CMU-TREE-TO-TREE	Carnegie Mellon University (Ammar et al., 2013)
CU-BOJAR, CU-DEPFIX, CU-TAMCHYNA	Charles University in Prague (Bojar et al., 2013)
CU-KAREL, CU-ZEMAN	Charles University in Prague (Bílek and Zeman, 2013)
CU-PHRASEFIX, CU-TECTOMT	Charles University in Prague (Galuščáková et al., 2013)
DCU	Dublin City University (Rubino et al., 2013a)
DCU-FDA	Dublin City University (Bicici, 2013a)
DCU-OKITA	Dublin City University (Okita et al., 2013)
DESRT	Università di Pisa (Miceli Barone and Attardi, 2013)
ITS-LATL	University of Geneva
JHU	Johns Hopkins University (Post et al., 2013)
KIT	Karlsruhe Institute of Technology (Cho et al., 2013)
LIA	Université d'Avignon (Huet et al., 2013)
LIMSI	LIMSI (Allauzen et al., 2013)
MES-*	Munich / Edinburgh / Stuttgart (Durrani et al., 2013a; Weller et al., 2013)
OMNIFLUENT	SAIC (Matusov and Leusch, 2013)
PROMT	PROMT Automated Translations Solutions
QCRI-MES	Qatar / Munich / Edinburgh / Stuttgart (Sajjad et al., 2013)
QUAERO	QUAERO (Peitz et al., 2013a)
RWTH	RWTH Aachen (Peitz et al., 2013b)
SHEF	University of Sheffield
STANFORD	Stanford University (Green et al., 2013)
TALP-UPC	TALP Research Centre (Formiga et al., 2013a)
TUBITAK	TÜBİTAK-BİLGEM (Durgar El-Kahlout and Mermer, 2013)
UCAM	University of Cambridge (Pino et al., 2013)
UEDIN, UEDIN-HEAFIELD	University of Edinburgh (Durrani et al., 2013b)
UEDIN-SYNTAX	University of Edinburgh (Nadejde et al., 2013)
UMD	University of Maryland (Eidelman et al., 2013)
UU	Uppsala University (Stymne et al., 2013)
COMMERCIAL-1,2,3	*Anonymized commercial systems*
ONLINE-A,B,G	*Anonymized online systems*

Table 1: 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.

3.1 Ranking translations of sentences

The ranking among systems is produced by collecting a large number of rankings between the systems' translations. Every language task had many participating systems (the largest was 19, for the Russian-English task). Rather than asking judges to provide a complete ordering over all the translations of a source segment, we instead randomly select five systems and ask the judge to rank just those. We call each of these a *ranking task*. A screenshot of the ranking interface is shown in Figure 2.

For each ranking task, the judge is presented with a source segment, a reference translation, and the outputs of five systems (anonymized and randomly-ordered). The following simple 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).*

The rankings of the systems are numbered from 1 to 5, with 1 being the best translation and 5 being the worst. Each ranking task has the potential to provide 10 *pairwise rankings*, and fewer if the judge chooses any ties. For example, the ranking

{A:1, B:2, C:4, D:3, E:5}

provides 10 pairwise rankings, while the ranking

{A:3, B:3, C:4, D:3, E:1}

provides just 7. The absolute value of the ranking or the degree of difference is not considered.

We use the collected pairwise rankings to assign each system a score that reflects how highly that system was usually ranked by the annotators. The score for some system A reflects how frequently it was judged to be better than other systems when compared on the same segment; its score is the number of pairwise rankings where it was judged to be better, divided by the total number of non-tying pairwise comparisons. These scores were used to compute clusters of systems and rankings between them (§3.4).

3.2 Collecting the data

A goal this year was to collect enough data to achieve statistical significance in the rankings. We distributed the workload among two groups of judges: *researchers* and *Turkers*. The researcher group comprised partipants in the shared task, who were asked to contribute judgments on 300 sentences for each system they contributed. The researcher evaluation was held over three weeks from May 17–June 7, and yielded about 280k pairwise rankings.

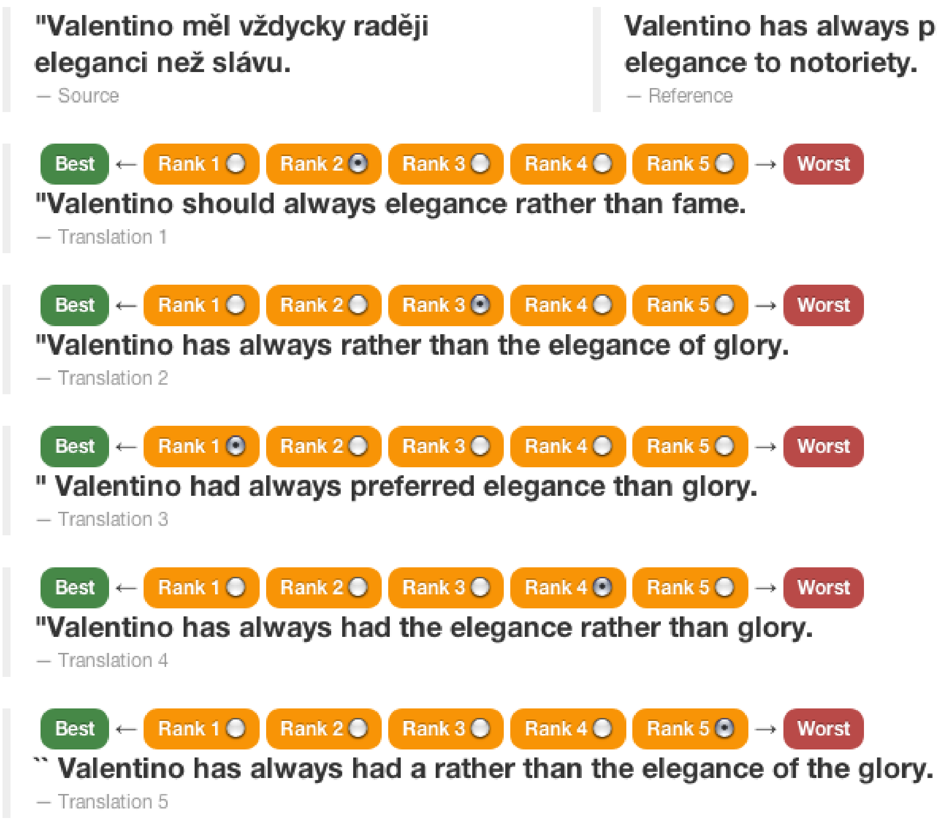
The Turker group was composed of non-expert annotators hired on Amazon's Mechanical Turk (MTurk). A basic unit of work on MTurk is called a Human Intelligence Task (HIT) and included three ranking tasks, for which we paid $0.25. To ensure that the Turkers provided high quality annotations, this portion of the evaluation was begun after the researcher portion had completed, enabling us to embed controls in the form of high-consensus pairwise rankings in the Turker HITs. To build these controls, we collected ranking tasks containing pairwise rankings with a high degree of researcher consensus. An example task is here:

SENTENCE	504
SOURCE	*Vor den heiligen Stätten verbeugen*
REFERENCE	*Let's worship the holy places*
SYSTEM A	Before the holy sites curtain
SYSTEM B	Before we bow to the Holy Places
SYSTEM C	To the holy sites bow
SYSTEM D	Bow down to the holy sites
SYSTEM E	Before the holy sites pay

		A	B	C	D	E
	A	-	0	0	0	3
	B	5	-	0	1	5
MATRIX	C	6	6	-	0	6
	D	6	8	5	-	6
	E	0	0	0	0	-

Matrix entry $M_{i,j}$ records the number of researchers who judged System i to be better than System j. We use as controls pairwise judgments for which $|M_{i,j} - M_{j,i}| > 5$, i.e., judgments where the researcher consensus ran strongly in one direction. We rejected HITs from Turkers who encountered at least 10 of these controls and failed more than 50% of them.

There were 463 people who participated in the Turker portion of the manual evaluation, contributing 664k pairwise rankings from Turkers who passed the controls. Together with the researcher judgments, we collected close to a million pairwise rankings, compared to 101k collected last year: a ten-fold increase. Table 2 contains more detail.



"Valentino měl vždycky raději eleganci než slávu.
— Source

Valentino has always preferred elegance to notoriety.
— Reference

Best ← Rank 1 ○ Rank 2 ● Rank 3 ○ Rank 4 ○ Rank 5 ○ → Worst
"Valentino should always elegance rather than fame.
— Translation 1

Best ← Rank 1 ○ Rank 2 ○ Rank 3 ● Rank 4 ○ Rank 5 ○ → Worst
"Valentino has always rather than the elegance of glory.
— Translation 2

Best ← Rank 1 ● Rank 2 ○ Rank 3 ○ Rank 4 ○ Rank 5 ○ → Worst
" Valentino had always preferred elegance than glory.
— Translation 3

Best ← Rank 1 ○ Rank 2 ○ Rank 3 ○ Rank 4 ● Rank 5 ○ → Worst
"Valentino has always had the elegance rather than glory.
— Translation 4

Best ← Rank 1 ○ Rank 2 ○ Rank 3 ○ Rank 4 ○ Rank 5 ● → Worst
`` Valentino has always had a rather than the elegance of the glory.
— Translation 5

Figure 2: Screenshot of the Appraise interface used in the human evaluation campaign. The annotator is presented with a source segment, a reference translation, and the outputs of five systems (anonymized and randomly-ordered) and has to rank these according to their translation quality, ties are allowed. For technical reasons, annotators on Amazon's Mechanical Turk received all three ranking tasks for a single HIT on a single page, one upon the other.

3.3 Annotator agreement

Each year we calculate annotator agreement scores for the human evaluation as a measure of the reliability of the rankings. We measured pairwise agreement among annotators using Cohen's kappa coefficient (κ) (Cohen, 1960), which is defined as

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

where $P(A)$ is the proportion of times that the annotators agree, and $P(E)$ is the proportion of time that they would agree by chance. 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 systems 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 should capture the probability that two annotators would agree randomly. Therefore:

$$P(E) = P(\textsc{a}{>}\textsc{b})^2 + P(\textsc{a}{=}\textsc{b})^2 + P(\textsc{a}{<}\textsc{b})^2$$

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 3 gives κ values for inter-annotator agreement for WMT11–WMT13 while Table 4 details intra-annotator agreement scores. Due to the change of annotation software, we used a slightly different way of computing annotator agreement scores. Therefore, we chose to re-compute values for previous WMTs to allow for a fair comparison. 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,

LANGUAGE PAIR	Systems	Rankings	Average
Czech-English	11	85,469	7,769.91
English-Czech	12	102,842	8,570.17
German-English	17	128,668	7,568.71
English-German	15	77,286	5,152.40
Spanish-English	12	67,832	5,652.67
English-Spanish	13	60,464	4,651.08
French-English	13	80,741	6,210.85
English-French	17	100,783	5,928.41
Russian-English	19	151,422	7,969.58
English-Russian	14	87,323	6,237.36
Total	148	942,840	6,370.54
WMT12	103	101,969	999.69
WMT11	133	63,045	474.02

Table 2: Amount of data collected in the WMT13 manual evaluation. The final two rows report summary information from the previous two workshops.

LANGUAGE PAIR	WMT11	WMT12	WMT13	WMT13$_r$	WMT13$_m$
Czech-English	0.400	0.311	0.244	0.342	0.279
English-Czech	0.460	0.359	0.168	0.408	0.075
German-English	0.324	0.385	0.299	0.443	0.324
English-German	0.378	0.356	0.267	0.457	0.239
Spanish-English	0.494	0.298	0.277	0.415	0.295
English-Spanish	0.367	0.254	0.206	0.333	0.249
French-English	0.402	0.272	0.275	0.405	0.321
English-French	0.406	0.296	0.231	0.434	0.237
Russian-English	—	—	0.278	0.315	0.324
English-Russian	—	—	0.243	0.416	0.207

Table 3: κ scores measuring inter-annotator agreement. The WMT13$_r$ and WMT13$_m$ columns provide breakdowns for researcher annotations and MTurk annotations, respectively. See Table 4 for corresponding intra-annotator agreement scores.

0.6–0.8 is substantial, and 0.8–1.0 is almost perfect. We find that the agreement rates are more or less the same as in prior years.

The WMT13 column contains both researcher and Turker annotations at a roughly 1:2 ratio. The final two columns break out agreement numbers between these two groups. The researcher agreement rates are similar to agreement rates from past years, while the Turker agreement are well below researcher agreement rates, varying widely, but often comparable to WMT11 and WMT12. Clearly, researchers are providing us with more consistent opinions, but whether these differences are explained by Turkers racing through jobs, the particularities that inform researchers judging systems they know well, or something else, is hard to tell. Intra-annotator agreement scores are also on par from last year's level, and are often much better. We observe better intra-annotator agreement for researchers compared to Turkers.

As a small test, we varied the threshold of acceptance against the controls for the Turker data alone and computed inter-annotator agreement scores on the datasets for the Russian–English task (the only language pair where we had enough data at high thresholds). Table 5 shows that higher thresholds do indeed give us better agreements, but not monotonically. The increasing κs suggests that we can find a segment of Turkers who do a better job and that perhaps a slightly higher threshold of 0.6 would serve us better, while the remaining difference against the researchers suggests there may be different mindsets informing the decisions. In any case, getting the best performance out of the Turkers remains difficult.

3.4 System Score

Given the multitude of pairwise comparisons, we would like to rank the systems according to a single score computed for each system. In re-

LANGUAGE PAIR	WMT11	WMT12	WMT13	WMT13$_r$	WMT13$_m$
Czech-English	0.597	0.454	0.479	0.483	0.478
English-Czech	0.601	0.390	0.290	0.547	0.242
German-English	0.576	0.392	0.535	0.643	0.515
English-German	0.528	0.433	0.498	0.649	0.452
Spanish-English	0.574	1.000	0.575	0.605	0.537
English-Spanish	0.426	0.329	0.492	0.468	0.492
French-English	0.673	0.360	0.578	0.585	0.565
English-French	0.524	0.414	0.495	0.630	0.486
Russian-English	—	—	0.450	0.363	0.477
English-Russian	—	—	0.513	0.582	0.500

Table 4: κ scores measuring intra-annotator agreement, i.e., self-consistency of judges, across for the past few years of the human evaluation. The WMT13$_r$ and WMT13$_m$ columns provide breakdowns for researcher annotations and MTurk annotations, respectively. The perfect inter-annotator agreement for Spanish-English is a result of there being very little data for that language pair.

thresh.	rankings	κ
0.5	16,605	0.234
0.6	9,999	0.337
0.7	3,219	0.360
0.8	1,851	0.395
0.9	849	0.336

Table 5: Agreement as a function of threshold for Turkers on the Russian–English task. The threshold is the percentage of controls a Turker must pass for her rankings to be accepted.

cent evaluation campaigns, we tweaked the metric and now arrived at a intuitive score that has been demonstrated to be accurate in ranking systems according to their true quality (Koehn, 2012).

The score, which we call EXPECTED WINS, has an intuitive explanation. If the system is compared against a randomly picked opposing system, on a randomly picked sentence, by a randomly picked judge, what is the probability that its translation is ranked higher?

Formally, the score for a system S_i among a set of systems $\{S_j\}$ given a pool of pairwise rankings summarized as win(A, B) — the number of times system A is ranked higher than system B — is defined as follows:

$$\text{score}(S_i) = \frac{1}{|\{S_j\}|} \sum_{j,j \neq i} \frac{\text{win}(S_i, S_j)}{\text{win}(S_i, S_j) + \text{win}(S_j, S_i)}$$

Note that this score ignores ties.

3.5 Rank Ranges and Clusters

Given the scores, we would like to rank the systems, which is straightforward. But we would also like to know, if the obtained system ranking is statistically significant. Typically, given the large number of systems that participate, and the similarity of the systems given a common training data condition and often common toolsets, there will be some systems that will be very close in quality.

To establish the reliability of the obtained system ranking, we use bootstrap resampling. We sample from the set of pairwise rankings an equal sized set of pairwise rankings (allowing for multiple drawings of the same pairwise ranking), compute the expected wins score for each system based on this sample, and rank each system. By repeating this procedure a 1,000 times, we can determine a range of ranks, into which system falls at least 95% of the time (i.e., at least 950 times) — corresponding to a p-level of $p \leq 0.05$.

Furthermore, given the rank ranges for each system, we can cluster systems with overlapping rank ranges.[8]

For all language pairs and all systems, Table 6 reports all system scores, rank ranges, and clusters. The official interpretation of these results is that systems in the same cluster are considered tied. Given the large number of judgements 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.

[8] Formally, given ranges defined by start(S_i) and end(S_i), we seek the largest set of clusters $\{C_c\}$ that satisfies:

$$\begin{aligned}
&\forall S\ \exists C : S \in C \\
&S \in C_a, S \in C_b \rightarrow C_a = C_b \\
&C_a \neq C_b \rightarrow \forall S_i \in C_a, S_j \in C_b : \\
&\qquad \text{start}(S_i) > \text{end}(S_j) \text{ or start}(S_j) > \text{end}(S_i)
\end{aligned}$$

Czech-English

#	score	range	system
1	0.607	1	UEDIN-HEAFIELD
2	0.582	2-3	ONLINE-B
	0.573	2-4	MES
	0.562	3-5	UEDIN
	0.547	4-7	ONLINE-A
	0.542	5-7	UEDIN-SYNTAX
	0.534	6-7	CU-ZEMAN
8	0.482	8	CU-TAMCHYNA
9	0.458	9	DCU-FDA
10	0.321	10	JHU
11	0.297	11	SHEF-WPROA

English-Czech

#	score	range	system
1	0.580	1-2	CU-BOJAR
	0.578	1-2	CU-DEPFIX
3	0.562	3	ONLINE-B
4	0.525	4	UEDIN
5	0.505	5-7	CU-ZEMAN
	0.502	5-7	MES
	0.499	5-8	ONLINE-A
	0.484	7-9	CU-PHRASEFIX
	0.476	8-9	CU-TECTOMT
10	0.457	10-11	COMMERCIAL-1
	0.450	10-11	COMMERCIAL-2
12	0.389	12	SHEF-WPROA

Spanish-English

#	score	range	system
1	0.624	1	UEDIN-HEAFIELD
2	0.595	2	ONLINE-B
3	0.570	3-5	UEDIN
	0.570	3-5	ONLINE-A
	0.567	3-5	MES
6	0.537	6	LIMSI-SOUL
7	0.514	7	DCU
8	0.488	8-9	DCU-OKITA
	0.484	8-9	DCU-FDA
10	0.462	10	CU-ZEMAN
11	0.425	11	JHU
12	0.169	12	SHEF-WPROA

English-Spanish

#	rank	range	system
1	0.637	1	ONLINE-B
2	0.582	2-4	ONLINE-A
	0.578	2-4	UEDIN
	0.567	3-4	PROMT
5	0.535	5-6	MES
	0.528	5-6	TALP-UPC
7	0.491	7-8	LIMSI
	0.474	7-9	DCU
	0.472	8-10	DCU-FDA
	0.455	9-11	DCU-OKITA
	0.446	10-11	CU-ZEMAN
12	0.417	12	JHU
13	0.324	13	SHEF-WPROA

German-English

#	rank	range	system
1	0.660	1	ONLINE-B
2	0.620	2-3	ONLINE-A
	0.608	2-3	UEDIN-SYNTAX
4	0.586	4-5	UEDIN
	0.584	4-5	QUAERO
	0.571	5-7	KIT
	0.562	6-7	MES
8	0.543	8-9	RWTH-JANE
	0.533	8-10	MES-REORDER
	0.526	9-10	LIMSI-SOUL
11	0.480	11	TUBITAK
12	0.462	12-13	UMD
	0.462	12-13	DCU
14	0.396	14	CU-ZEMAN
15	0.367	15	JHU
16	0.311	16	SHEF-WPROA
17	0.238	17	DESRT

English-German

#	rank	range	system
1	0.637	1-2	ONLINE-B
	0.636	1-2	PROMT
3	0.614	3	UEDIN-SYNTAX
	0.587	3-5	ONLINE-A
	0.571	4-6	UEDIN
	0.554	5-6	KIT
7	0.523	7	STANFORD
8	0.507	8	LIMSI-SOUL
9	0.477	9-11	MES-REORDER
	0.476	9-11	JHU
	0.460	10-12	CU-ZEMAN
	0.453	11-12	TUBITAK
13	0.361	13	UU
14	0.329	14-15	SHEF-WPROA
	0.323	14-15	RWTH-JANE

English-Russian

#	rank	range	system
1	0.641	1	PROMT
2	0.623	2	ONLINE-B
3	0.556	3-4	CMU
	0.542	3-6	ONLINE-G
	0.538	3-7	ONLINE-A
	0.531	4-7	UEDIN
	0.520	5-7	QCRI-MES
8	0.498	8	CU-KAREL
9	0.478	9-10	MES-QCRI
	0.469	9-10	JHU
11	0.434	11-12	COMMERCIAL-3
	0.426	11-13	LIA
	0.419	12-13	BALAGUR
14	0.331	14	CU-ZEMAN

French-English

#	rank	range	system
1	0.638	1	UEDIN-HEAFIELD
2	0.604	2-3	UEDIN
	0.591	2-3	ONLINE-B
4	0.573	4-5	LIMSI-SOUL
	0.562	4-5	KIT
	0.541	5-6	ONLINE-A
7	0.512	7	MES-SIMPLIFIED
8	0.486	8	DCU
9	0.439	9-10	RWTH
	0.429	9-11	CMU-T2T
	0.420	10-11	CU-ZEMAN
12	0.389	12	JHU
13	0.322	13	SHEF-WPROA

English-French

#	rank	range	system
1	0.607	1-2	UEDIN
	0.600	1-3	ONLINE-B
	0.588	2-4	LIMSI-SOUL
	0.584	3-4	KIT
5	0.553	5-7	PROMT
	0.551	5-8	STANFORD
	0.547	5-8	MES
	0.537	6-9	MES-INFLECTION
	0.533	7-10	RWTH-PB
	0.516	9-11	ONLINE-A
	0.499	10-11	DCU
12	0.427	12	CU-ZEMAN
13	0.408	13	JHU
14	0.382	14	OMNIFLUENT
15	0.350	15	ITS-LATL
16	0.326	16	ITS-LATL-PE

Russian-English

#	rank	range	system
1	0.657	1	ONLINE-B
2	0.604	2-3	CMU
	0.588	2-3	ONLINE-A
4	0.562	4-6	ONLINE-G
	0.561	4-6	PROMT
	0.550	5-7	QCRI-MES
	0.546	5-7	UCAM
8	0.527	8-9	BALAGUR
	0.519	8-10	MES-QCRI
	0.507	9-11	UEDIN
	0.497	10-12	OMNIFLUENT
	0.492	11-14	LIA
	0.483	12-15	OMNIFLUENT-C
	0.481	12-15	UMD
	0.476	13-15	CU-KAREL
16	0.432	16	COMMERCIAL-3
17	0.417	17	UEDIN-SYNTAX
18	0.396	18	JHU
19	0.215	19	CU-ZEMAN

Table 6: Official results for the WMT13 translation task. Systems are ordered by the expected win score. Lines between systems indicate clusters according to bootstrap resampling at p-level $p \leq .05$. This method is also used to determine the range of ranks into which system falls. Systems with grey background indicate use of resources that fall outside the constraints provided for the shared task.

4 Understandability of English→Czech

For the English-to-Czech translation, we conducted a variation of the "understandability" test as introduced in WMT09 (Callison-Burch et al., 2009) and used in WMT10. In order to obtain additional reference translations, we conflated this test with post-editing. The procedure was as follows:

1. **Monolingual editing** (also called blind editing). The first annotator is given just the MT output and requested to correct it. Given errors in MT outputs, some guessing of the original meaning is often inevitable and the annotators are welcome to try. If unable, they can mark the sentences as incomprehensible.
2. **Review**. A second annotator is asked to validate the monolingual edit given both the source and reference translations. Our instructions specify three options:
   (a) If the monolingual edit is an adequate translation and acceptably fluent Czech, confirm it without changes.
   (b) If the monolingual edit is adequate but needs polishing, modify the sentence and prefix it with the label 'OK:'.
   (c) If the monolingual edit is wrong, correct it. You may start from the original unedited MT output, if that is easier. Avoid using the reference directly, prefer words from MT output whenever possible.

The motivation behind this procedure is that we want to save the time necessary for reading the sentence. If the reviewer has already considered whether the sentence is an acceptable translation, they do not need to read the MT output again in order to post-edit it. Our approach is thus somewhat the converse of Aziz et al. (2013) who analyze post-editing effort to obtain rankings of MT systems. We want to measure the understandability of MT outputs and obtain post-edits at the same time.

Both annotation steps were carried out in the CASMACAT/Matecat post-editing user interface.[9], modified to provide the relevant variants of the sentence next to the main edit box. Screenshots of the two annotation phases are given in Figure 3 and Figure 4.

Occurrence	GOOD	ALMOST	BAD	EMPTY	Total
First	34.7	0.1	42.3	11.0	4082
Repeated	41.1	0.1	41.0	6.1	805
Overall	35.8	0.1	42.1	10.2	4887

Table 7: Distribution of review statuses.

Similarly to the traditional ranking task, we provided three consecutive sentences from the original text, each translated with a different MT system. The annotators are free to use this contextual information when guessing the meaning or reviewing the monolingual edits. Each "annotation HIT" consists of 24 sentences, i.e. 8 snippets of 3 consecutive sentences.

4.1 Basic Statistics on Editing

In total, 21 annotators took part in the exercise, 20 of them contributed to monolingual editing and 19 contributed to the reviews.

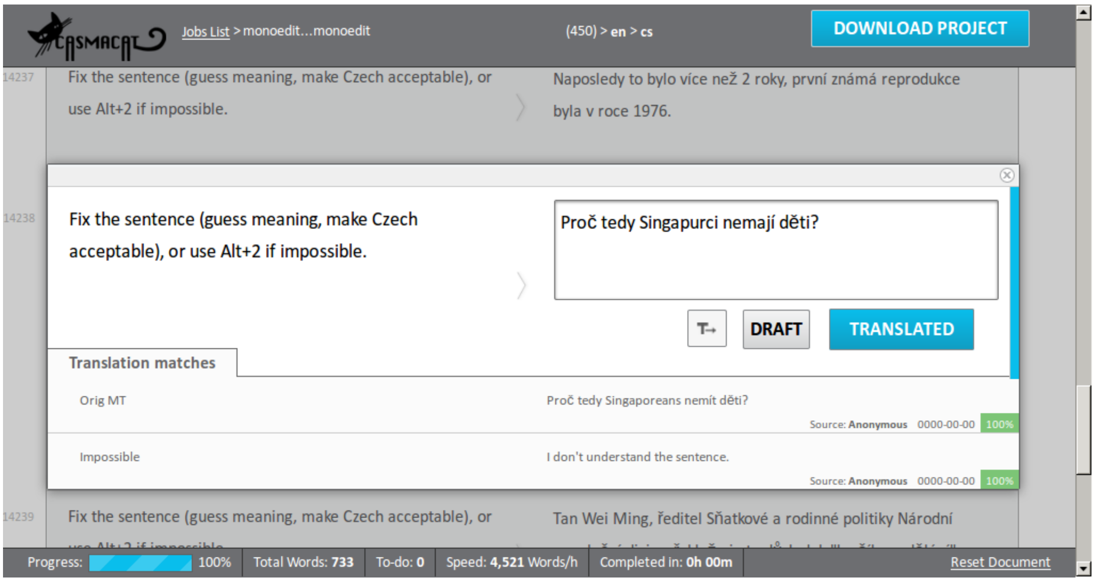
Connecting each review with the monolingual edit (some edits received multiple reviews), we obtain one data row. We collected 4887 data rows (i.e. sentence revisions) for 3538 monolingual edits, covering 1468 source sentences as translated by 12 MT systems (including the reference).

Not all MT systems were considered for each sentence, we preferred to obtain judgments for more source sentences.

Based on the annotation instructions, each data row has one of the four possible statuses: GOOD, ALMOST, BAD, and EMPTY. GOOD rows are those where the reviewer accepted the monolingual edit without changes, ALMOST edits were modified by the reviewer but they were marked as 'OK'. BAD edits were changed by the reviewer and no 'OK' mark was given. Finally, the status EMPTY is assigned to rows where the monolingual editor refused to edit the sentence. The EMPTY rows nevertheless contain the ("regular") post-edit of the reviewer, so they still provide a new reference translation for the sentence.

Table 7 summarizes the distribution of row statuses depending on one more significant distinction: whether the monolingual editor has seen the sentence before or not. We see that EMPTY and BAD monolingual edits together drop by about 6% absolute when the sentence is not new to the monolingual editor. The occurrence is counted as "repeated" regardless whether the annotator has previously seen the sentence in an editing or reviewing task. Unless stated otherwise, we exclude repeated edits from our calculations.

[9] http://www.casmacat.eu/index.php?n=Workbench



14237 Fix the sentence (guess meaning, make Czech acceptable), or use Alt+2 if impossible.

Naposledy to bylo více než 2 roky, první známá reprodukce byla v roce 1976.

14238 Fix the sentence (guess meaning, make Czech acceptable), or use Alt+2 if impossible.

Proč tedy Singapurci nemají děti?

Orig MT	Proč tedy Singaporeans nemít děti?	Source: Anonymous 0000-00-00 100%
Impossible	I don't understand the sentence.	Source: Anonymous 0000-00-00 100%

14239 Fix the sentence (guess meaning, make Czech acceptable), or use Alt+2 if impossible

Tan Wei Ming, ředitel Sňatkové a rodinné politiky Národní

Progress: 100% | Total Words: 733 | To-do: 0 | Speed: 4,521 Words/h | Completed in: 0h 00m

Figure 3: In this screen, the annotator is expected to correct the MT output given only the context of at most two neighbouring machine-translated sentences.

	ALMOST	Pairwise		
	treated	Comparisons	Agreement	κ
inter	separate	2690	56.0	0.270
	as BAD	2690	67.9	0.351
	as GOOD	2690	65.2	0.289
intra	separate	170	65.3	0.410
	as BAD	170	69.4	0.386
	as GOOD	170	71.8	0.422

Table 8: Annotator agreement when reviewing monolingual edits.

4.2 Agreement on Understandability

Before looking at individual system results, we consider annotator agreement in the review step. Details are given in Table 8. Given a (non-EMPTY) string from a monolingual edit, we would like to know how often two acceptability judgments by two different reviewers (inter-) or the same reviewer (intra-) agree. The repeated edits remain in this analysis because we are not interested in the origin of the string.

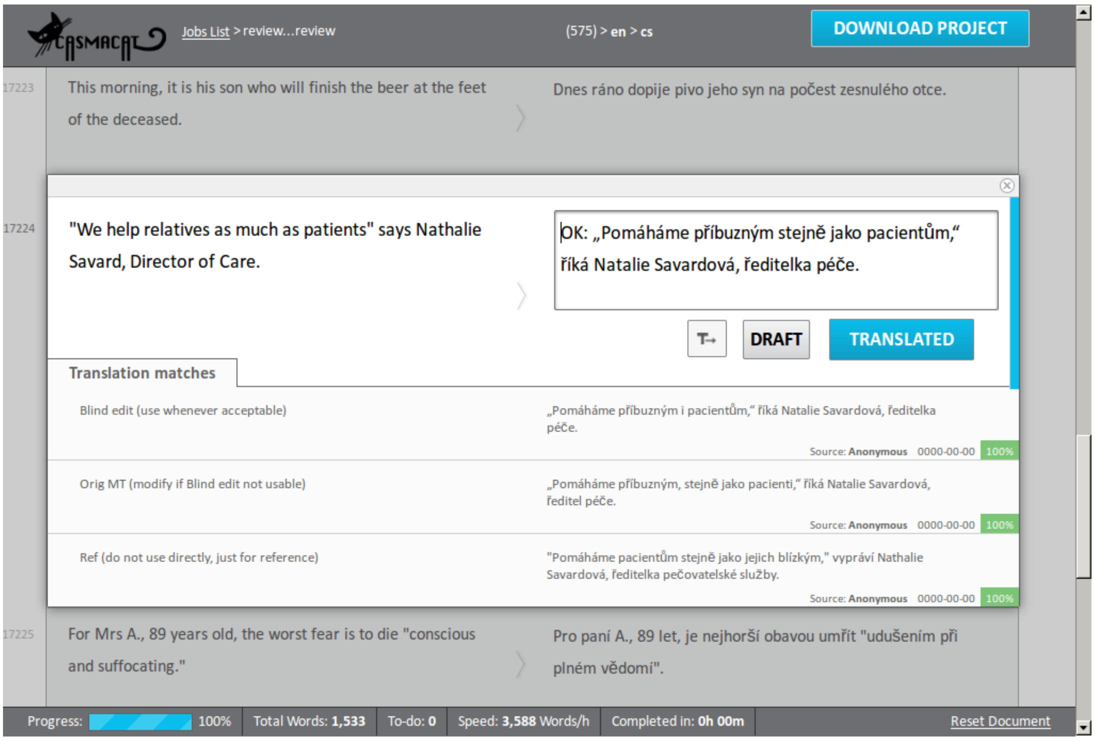
Our annotation setup leads to three possible labels: GOOD, ALMOST, and BAD. The agreement on one of three classes is bound to be lower than the agreement on two classes, so we also re-interpret ALMOST as either GOOD or BAD. Generally speaking, ALMOST is a positive judgment, so it would be natural to treat it as GOOD. However, in our particular setup, when the reviewer modified the sentence and *forgot* to add the label 'OK:', the item ended up in the BAD class. We conclude that this is indeed the case: the inter-annotator agreement appears higher if ALMOST is treated as BAD. Future versions of the reviewing interface should perhaps first ask for the yes/no judgment and only then allow to post-edit.

The κ values in Table 8 are the Fleiss' kappa (Fleiss, 1971), accounting for agreement by chance given the observed label distributions.

In WMT09, the agreements for this task were higher: 77.4 for inter-AA and 86.6 for intra-AA. (In 2010, the agreements for this task were not reported.) It is difficult to say whether the difference lies in the particular language pair, the different set of annotators, or the different user interface for our reviewing task. In 2009 and 2010, the reviewers were shown 5 monolingual edits at once and they were asked to judge each as acceptable or not acceptable. We show just one segment and they have probably set their minds on the post-editing rather than acceptability judgment. We believe that higher agreements can be reached if the reviewers first validate one or more of the edits and only then are allowed to post-edit it.

4.3 Understandability of English→Czech

Table 9 brings about the first main result of our post-editing effort. For each system (including the reference translation), we check how often a monolingual edit was marked OK or ALMOST by the subsequent reviewer. The average understandability across all MT systems into Czech is 44.2±1.6%. This is a considerable improvement compared to 2009 where the best systems produced about 32% understandable sentences. In



17223 This morning, it is his son who will finish the beer at the feet of the deceased.

Dnes ráno dopije pivo jeho syn na počest zesnulého otce.

17224 "We help relatives as much as patients" says Nathalie Savard, Director of Care.

OK: „Pomáháme příbuzným stejně jako pacientům,“ říká Natalie Savardová, ředitelka péče.

Blind edit (use whenever acceptable) — „Pomáháme příbuzným i pacientům,“ říká Natalie Savardová, ředitelka péče. Source: Anonymous 0000-00-00 100%

Orig MT (modify if Blind edit not usable) — „Pomáháme příbuzným, stejně jako pacienti,“ říká Natalie Savardová, ředitel péče. Source: Anonymous 0000-00-00 100%

Ref (do not use directly, just for reference) — "Pomáháme pacientům stejně jako jejich blízkým," vypráví Nathalie Savardová, ředitelka pečovatelské služby. Source: Anonymous 0000-00-00 100%

17225 For Mrs A., 89 years old, the worst fear is to die "conscious and suffocating."

Pro paní A., 89 let, je nejhorší obavou umřít "udušením při plném vědomí".

Figure 4: In this screen, the annotator is expected to validate the monolingual edit, correcting it if necessary. The annotator is expected to add the prefix ‘OK:’ if the correction was more or less cosmetic.

Rank	System	Total Observations	% Understandable
	Overall incl. ref.	4082	46.7±1.6
	Overall without ref.	3808	44.2±1.6
1	Reference	274±31	80.3±4.8
2-6	CU-ZEMAN	348±34	51.7±5.1
2-6	UEDIN	332±33	51.5±5.4
2-6	ONLINE-B	337±34	50.7±5.3
2-6	CU-BOJAR	341±35	50.7±5.2
2-7	CU-DEPFIX	350±34	48.0±5.3
6-10	COMMERCIAL-2	358±36	43.6±5.2
6-11	COMMERCIAL-1	316±34	41.5±5.5
7-12	CU-TECTOMT	338±34	39.4±5.2
8-12	MES	346±36	38.4±5.2
8-12	CU-PHRASEFIX	394±40	38.1±4.8
10-12	SHEF-WPROA	348±32	34.2±5.1
	2009 Reference		91
	2009 Best System		32
	2010 Reference		97
	2010 Best System		58

Table 9: Understandability of English→Czech systems. The ± values indicate empirical confidence bounds at 95%. Rank ranges were also obtained in the same resampling: in 95% of observations, the system was ranked in the given range.

2010, the best systems or system combinations reached 55%–58%. The test set across years and the quality of references and judgments also play a role. In our annotation setup, the references appear to be correctly understandable only to 80.3±4.8%.

To estimate the variance of these results due to the particular sentences chosen, we draw 1000 random samples from the dataset, preserving the dataset size and repeating some. The exact number of judgments per system can thus vary. We report the 95% empirical confidence interval after the ‘±’ signs in Table 9 (the systems range from ±4.8 to ±5.5). When we drop individual blind editors or reviewers, the understandability judgments differ by about ±2 to ±4. In other words, the dependence on the test set appears higher than the dependence on the annotators.

The limited size of our dataset allows us only to separate two main groups of systems: those ranking 2–6 and those ranking worse. This rough grouping vaguely matches with WMT13 ranking results as given in Table 6. A somewhat surprising observation is that two automatic corrections ranked better in WMT13 ranking but score worse in understandability: CU-DEPFIX fixes some lost negation and some agreement errors of CU-BOJAR and CU-PHRASEFIX is a standard statistical post-editing of a transfer-based system CU-TECTOMT. A detailed inspection of the data is necessary to explain this.

5 More Reference Translations for Czech

Our annotation procedure described in Section 4 allowed us to obtain a considerable number of additional reference translations on top of official single reference.

Refs	1	2	3	4	5	6	7	8	9	10-16
Sents	233	709	174	123	60	48	40	27	25	29

Table 10: Number of source sentences with the given number of distinct reference translations.

In total, our edits cover 1468 source sentences, i.e. about a half of the official test set size, and provide 4311 unique references. On average, one sentence in our set has 2.94±2.17 unique reference translations. Table 10 provides a histogram.

It is well known that automatic MT evaluation methods perform better with more references, because a single one may not confirm a correct part of MT output. This issue is more severe for morphologically rich languages like Czech where about 1/3 of MT output was correct but not confirmed by the reference (Bojar et al., 2010). Advanced evaluation methods apply paraphrasing to smooth out some of the lexical divergence (Kauchak and Barzilay, 2006; Snover et al., 2009; Denkowski and Lavie, 2010). Simpler techniques such as lemmatizing are effective for morphologically rich languages (Tantug et al., 2008; Kos and Bojar, 2009) but they will lose resolution once the systems start performing generally well.

WMTs have taken the stance that a big enough test set with just a single reference should compensate for the lack of other references. We use our post-edited reference translations to check this assumption for BLEU and NIST as implemented in `mteval-13a` (international tokenization switched on, which is not the default setting).

We run many probes, randomly picking the test set size (number of distinct sentences) and the number of distinct references per sentence. Note that such test sets are somewhat artificially more diverse; in narrow domains, source sentences can repeat and even appear verbatim in the training data, and in natural test sets with multiple references, short sentences can receive several identical translations.

For each probe, we measure the Spearman's rank correlation coefficient ρ of the ranks proposed by BLEU or NIST and the manual ranks. We use the same implementation as applied in the WMT13 Shared Metrics Task (Macháček and Bojar, 2013). Note that the WMT13 metrics task still uses the WMT12 evaluation method ignoring ties, not the expected wins. As Koehn (2012) shows, the two methods do not differ much.

Overall, the correlation is strongly impacted by the particular choice of test sentences and reference translations. By picking sentences randomly, similarly or equally sized test sets can reach different correlations. Indeed, e.g. for a test set of about 1500 distinct sentences selected from the 3000-sentence official test set (1 reference translation), we obtain correlations for BLEU between 0.86 and 0.94.

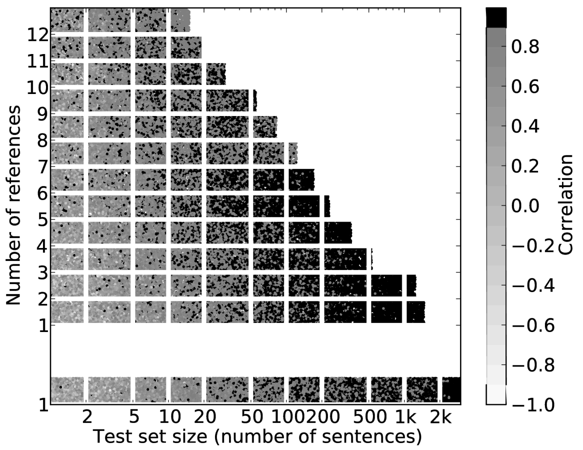


Figure 5: Correlation of BLEU and WMT13 manual ranks for English→Czech translation

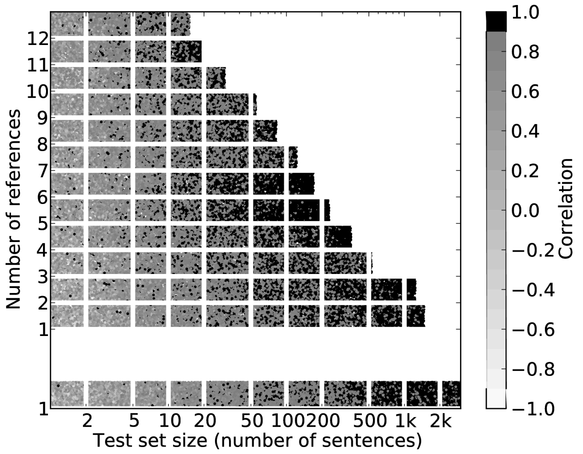


Figure 6: Correlation of NIST and WMT13 manual ranks for English→Czech translation

Figure 5 plots the correlations of BLEU and the system rankings, Figure 6 provides the same picture for NIST. The upper triangular part of the plot contains samples from our post-edited reference translations, the lower rectangular part contains probes from the official test set of 3000 sentences with 1 reference translation.

To interpret the observations, we also calculate the average and standard deviation of correlations for each cell in Figures 5 and 6. Figures 7 and 8 plot the values for 1, 6, 7 and 8 references for

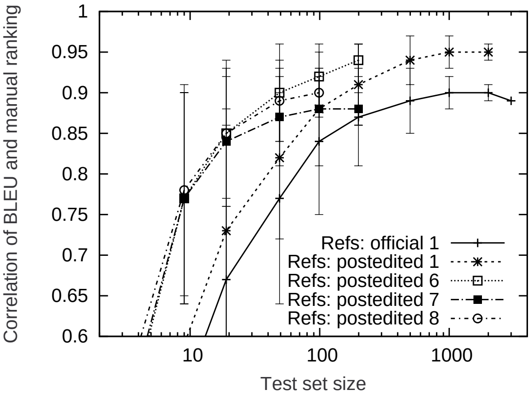


Figure 7: Projections from Figure 5 of BLEU and WMT13 manual ranks for English→Czech translation

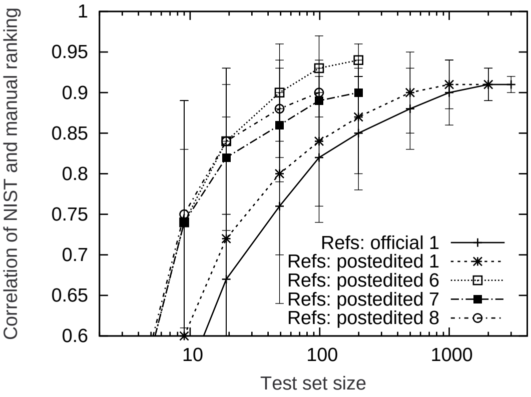


Figure 8: Projections from Figure 6 of NIST and WMT13 manual ranks for English→Czech translation

BLEU and NIST, resp. The projections confirm that the average correlations grow with test set size, the growth is however sub-logarithmic.

Starting from as few as a dozen of sentences, we see that using more references is better than using a larger test set. For BLEU, we however already seem to reach false positives at 7 references for one or two hundred sentences: larger sets with just one reference may correlate slightly better.

Using one reference obtained by post-editing seems better than using the official (independent) reference translations. BLEU is more affected than NIST by this difference even at relatively large test set size. Note that our post-edits are inspired by all MT systems, the good as well as the bad ones. This probably provides our set with a certain balance.

Overall, the best balance between the test set size and the number of references seems to lie somewhere around 7 references and 100 or 200 sentences. Creating such a test set could be even cheaper than the standard 3000 sentences with just one reference. However, the wide error bars remind us that even this setting can lead to correlations anywhere between 0.86 and 0.96. For other languages, data sets types or other MT evaluation methods, the best setting can be quite different and has to be sought for.

6 Quality Estimation Task

Machine translation quality estimation is the task of predicting a quality score for a machine translated text without access to reference translations. The most common approach is to treat the problem as a supervised machine learning task, using standard regression or classification algorithms. The second edition of the WMT shared task on quality estimation builds on the previous edition of the task (Callison-Burch et al., 2012), with variants to this previous task, including both sentence-level and word-level estimation, with new training and test datasets, along with evaluation metrics and baseline systems.

The motivation to include both sentence- and word-level estimation come from the different potential applications of these variants. Some interesting uses of sentence-level quality estimation are the following:

- Decide whether a given translation is good enough for publishing as is.
- Inform readers of the target language only whether or not they can rely on a translation.
- Filter out sentences that are not good enough for post-editing by professional translators.
- Select the best translation among options from multiple MT and/or translation memory systems.

Some interesting uses of word-level quality estimation are the following:

- Highlight words that need editing in post-editing tasks.
- Inform readers of portions of the sentence which are not reliable.
- Select the best segments among options from multiple translation systems for MT system combination.

The goals of this year's shared task were:

- To explore various granularity levels for the task (sentence-level and word-level).
- To explore the prediction of more objective scores such as edit distance and post-editing time.
- To explore the use of quality estimation techniques to replace reference-based MT evaluation metrics in the task of ranking alternative translations generated by different MT systems.
- To identify new and effective quality indicators (features) for all variants of the quality estimation task.
- To identify effective machine learning techniques for all variants of the quality estimation task.
- To establish the state of the art performance in the field.

Four subtasks were proposed, as we discuss in Sections 6.1 and 6.2. Each subtask provides specific datasets, annotated for quality according to the subtask (Section 6.3), and evaluates the system submissions using specific metrics (Section 6.6). When available, external resources (e.g. SMT training corpus) and translation engine-related resources were given to participants (Section 6.4), who could also use any additional external resources (no distinction between *open* and *close* tracks is made). Participants were also provided with a software package to extract quality estimation features and perform model learning (Section 6.5), with a suggested list of *baseline* features and learning method (Section 6.7). Participants could submit up to two systems for each subtask.

6.1 Sentence-level Quality Estimation

Task 1.1 Predicting Post-editing Distance This task is similar to the quality estimation task in WMT12, but with one important difference in the scoring variant: instead of using the post-editing effort scores in the [1-5] range, we use HTER (Snover et al., 2006) as quality score. This score is to be interpreted as the minimum edit distance between the machine translation and its manually post-edited version, and its range is [0, 1] (0 when no edit needs to be made, and 1 when all words need to be edited). Two variants of the results could be submitted in the shared task:

- **Scoring**: A quality score for each sentence translation in [0,1], to be interpreted as an HTER score; lower scores mean better translations.
- **Ranking**: A ranking of sentence translations for all source test sentences from best to worst. For this variant, it does not matter how the ranking is produced (from HTER predictions, likert predictions, or even without machine learning). The reference ranking is defined based on the true HTER scores.

Task 1.2 Selecting Best Translation This task consists in ranking up to five alternative translations for the same source sentence produced by multiple MT systems. We use essentially the same data provided to participants of previous years WMT's evaluation metrics task – where MT evaluation metrics are assessed according to how well they correlate with human rankings. However, reference translations produced by humans are not be used in this task.

Task 1.3 Predicting Post-editing Time For this task systems are required to produce, for each translation, the expected time (in seconds) it would take a translator to post-edit such an MT output. The main application for predictions of this type is in computer-aided translation where the predicted time can be used to select among different hypotheses or even to omit any MT output in cases where no good suggestion is available.

6.2 Word-level Quality Estimation

Based on the data of Task 1.3, we define Task 2, a word-level annotation task for which participants are asked to produce a label for each token that indicates whether the word should be changed by a post-editor or kept in the final translation. We consider the following two sets of labels for prediction:

- **Binary classification**: a keep/change label, the latter meaning that the token should be corrected in the post-editing process.
- **Multi-class classification**: a label specifying the edit action that should be performed on the token (keep as is, delete, or substitute).

6.3 Datasets

Task 1.1 Predicting post-editing distance For the training of models, we provided the WMT12 quality estimation dataset: 2,254 English-Spanish news sentences extracted from previous WMT translation task English-Spanish test sets (WMT09, WMT10, and WMT12). These were translated by a phrase-based SMT Moses system trained on Europarl and News Commentaries corpora as provided by WMT, along with their source sentences, reference translations, post-edited translations, and HTER scores. We used TERp (default settings: tokenised, case insensitive, etc., but capped to 1)[10] to compute the HTER scores. Likert scores in [1,5] were also provided, as participants may choose to use them for the ranking variant.

As test data, we use a subset of the WMT13 English-Spanish news test set with 500 sentences, whose translations were produced by the same SMT system used for the training set. To compute the true HTER labels, the translations were post-edited under the same conditions as those on the training set. As in any blind shared task, the HTER scores were solely used to evaluate the submissions, and were only released to participants after they submitted their systems.

A few variations of the training and test data were provided, including a version with cases restored and a version detokenized. In addition, we provided a number of engine-internal information from Moses for glass-box feature extraction, such as phrase and word alignments, model scores, word graph, n-best lists and information from the decoder's search graph.

Task 1.2 Selecting best translation As training data, we provided a large set of up to five alternative machine translations produced by different MT systems for each source sentence and ranked for quality by humans. This was the outcome of the manual evaluation of the translation task from WMT09-WMT12. It includes two language pairs: German-English and English-Spanish, with 7,098 and 4,592 source sentences and up to five ranked translations, totalling 32,922 and 22,447 translations, respectively.

As test data, a set of up to five alternative machine translations per source sentence from the WMT08 test sets was provided, with 365 (1,810) and 264 (1,315) source sentences (translations) for German-English and English-Spanish, respectively. We note that there was some overlap between the MT systems used in the training data and test datasets, but not all systems were the same, as different systems participate in WMT over the years.

Task 1.3 and Task 2 Predicting post-editing time and word-level edits For Tasks 1.3 and 2 we provides a new dataset consisting of 22 English news articles which were translated into Spanish using Moses and post-edited during a CASMACAT[11] field trial. Of these, 15 documents have been processed repeatedly by at least 2 out of 5 translators, resulting in a total of 1,087 segments. For each segment we provided:

- English source and Spanish translation.
- Spanish MT output which was used as basis for post-editing.
- Document and translator ID.
- Position of the segment within the document.

The metadata about translator and document was made available as we expect that translator performance and normalisation over document complexity can be helpful when predicting the time spend on a given segment.

For the training portion of the data we also provided:

- Time to post-edit in seconds (Task 1.3).
- Binary (Keep, Change) and multiclass (Keep, Substitute, Delete) labels on word level along with explicit tokenization (Task 2).

The labels in Task 2 are derived by computing WER between the original machine translation and its post-edited version.

6.4 Resources

For all tasks, we provided resources to extract quality estimation features when these were available:

- The SMT training corpus (WMT News and Europarl): source and target sides of the corpus used to train the SMT engines for Tasks 1.1, 1.3, and 2, and truecase models generated from these. These corpora can also be used for Task 1.2, but we note that some of the MT systems used in the datasets of this task were not statistical or did not use (only) the training corpus provided by WMT.

[10] http://www.umiacs.umd.edu/~snover/terp/

[11] http://casmacat.eu/

- Language models: n-gram language models of source and target languages generated using the SMT training corpora and standard toolkits such as SRILM Stolcke (2002), and a language model of POS tags for the target language. We also provided unigram, bigram and trigram counts.
- IBM Model 1 lexical tables generated by GIZA++ using the SMT training corpora.
- Phrase tables with word alignment information generated by scripts provided by Moses from the parallel corpora.
- For Tasks 1.1, 1.3 and 2, the Moses configuration file used for decoding or the code to re-run the entire Moses system.
- For Task 1.1, both English and Spanish resources for a number of advanced features such as pre-generated PCFG parsing models, topic models, global lexicon models and mutual information trigger models.

We refer the reader to the QuEst website[12] for a detailed list of resources provided for each task.

6.5 QuEst Framework

QuEst (Specia et al., 2013) is an open source framework for quality estimation which provides a wide variety of feature extractors from source and translation texts and external resources and tools. These range from simple, language-independent features, to advanced, linguistically motivated features. They include features that rely on information from the MT system that generated the translations (glass-box features), and features that are oblivious to the way translations were produced (black-box features).

QuEst also integrates a well-known machine learning toolkit, `scikit-learn`,[13] and other algorithms that are known to perform well on this task (e.g. Gaussian Processes), providing a simple and effective way of experimenting with techniques for feature selection and model building, as well as parameter optimisation through grid search.

From QuEst, a subset of 17 features and an SVM regression implementation were used as baseline for Tasks 1.1, 1.2 and 1.3. The software was made available to all participants.

6.6 Evaluation Metrics

Task 1.1 Predicting post-editing distance Evaluation is performed against the HTER and/or ranking of translations using the same metrics as in WMT12. For the **scoring** variant of the task, we use two standard metrics for regression tasks: Mean Absolute Error (MAE) as a primary metric, and Root of Mean Squared Error (RMSE) as a secondary metric. To improve readability, we report these error numbers by first mapping the HTER values to the $[0, 100]$ interval, to be read as percentage-points of the HTER metric. For a given test set S with entries $s_i, 1 \leq i \leq |S|$, we denote by $H(s_i)$ the proposed score for entry s_i (hypothesis), and by $V(s_i)$ the reference value for entry s_i (gold-standard value):

$$\text{MAE} = \frac{\sum_{i=1}^{N} |H(s_i) - V(s_i)|}{|S|}$$

$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^{N} (H(s_i) - V(s_i))^2}{|S|}}$$

Both these metrics are non-parametric, automatic and deterministic (and therefore consistent), and extrinsically interpretable. For instance, a MAE value of 10 means that, on average, the absolute difference between the hypothesized score and the reference score value is 10 percentage points (i.e., 0.10 difference in HTER scores). The interpretation of RMSE is similar, with the difference that RMSE penalises larger errors more (via the square function).

For the **ranking** variant of the task, we use the DeltaAvg metric proposed in the 2012 edition of the task (Callison-Burch et al., 2012) as our main metric. This metric assumes that each reference test instance has an extrinsic number associated with it that represents its ranking with respect to the other test instances. For completeness, we present here again the definition of DeltaAvg.

The goal of the DeltaAvg metric is to measure how valuable a proposed ranking (which we call a *hypothesis* ranking) is, according to the true ranking values associated with the test instances. We first define a parametrised version of this metric, called DeltaAvg$[n]$. The following notations are used: for a given entry sentence s, $V(s)$ represents the function that associates an extrinsic value to that entry; we extend this notation to a set S, with $V(S)$ representing the average of all $V(s), s \in S$.

[12] http://www.quest.dcs.shef.ac.uk/

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

Intuitively, $V(S)$ is a quantitative measure of the "quality" of the set S, as induced by the extrinsic values associated with the entries in S. For a set of ranked entries S and a parameter n, we denote by S_1 the first quantile of set S (the highest-ranked entries), S_2 the second quantile, and so on, for n quantiles of equal sizes.[14] We also use the notation $S_{i,j} = \bigcup_{k=i}^{j} S_k$. Using these notations, we define:

$$\text{DeltaAvg}_V[n] = \frac{\sum_{k=1}^{n-1} V(S_{1,k})}{n-1} - V(S)$$

When the valuation function V is clear from the context, we write DeltaAvg$[n]$ for DeltaAvg$_V[n]$. The parameter n represents the number of quantiles we want to split the set S into. For instance, $n = 2$ gives DeltaAvg[2] = $V(S_1)-V(S)$, hence it measures the difference between the quality of the top quantile (top half) S_1 and the overall quality (represented by $V(S)$). For $n = 3$, DeltaAvg[3] = $(V(S_1)+V(S_{1,2})/2-V(S) = ((V(S_1)-V(S))+(V(S_{1,2}-V(S)))/2$, hence it measures an average difference across two cases: between the quality of the top quantile (top third) and the overall quality, and between the quality of the top two quantiles ($S_1 \cup S_2$, top two-thirds) and the overall quality. In general, DeltaAvg$[n]$ measures an average difference in quality across $n-1$ cases, with each case measuring the impact in quality of adding an additional quantile, from top to bottom. Finally, we define:

$$\text{DeltaAvg}_V = \frac{\sum_{n=2}^{N} \text{DeltaAvg}_V[n]}{N-1}$$

where $N = |S|/2$. As before, we write DeltaAvg for DeltaAvg$_V$ when the valuation function V is clear from the context. The DeltaAvg metric is an average across all DeltaAvg$[n]$ values, for those n values for which the resulting quantiles have at least 2 entries (no singleton quantiles).

We present results for DeltaAvg using as valuation function V the HTER scores, as defined in Section 6.3. We also use Spearman's rank correlation coefficient ρ as a secondary metric.

Task 1.2 Selecting best translation The performance on the task of selecting the best translation from a pool of translation candidates is measured by comparing proposed (hypothesis) rankings against human-produced rankings. The metric used is Kendall's τ rank correlation coefficient, computed as follows:

$$\tau = \frac{|\text{concordant pairs}| - |\text{discordant pairs}|}{|\text{total pairs}|}$$

where a concordant pair is a pair of two translations for the same source segment in which the ranking order proposed by a human annotator and the ranking order of the hypothesis agree; in a discordant pair, they disagree. The possible values of τ range between 1 (where all pairs are concordant) and -1 (where all pairs are discordant). Thus a system with ranking predictions having a higher τ value makes predictions that are more similar to human judgements than a system with ranking predictions having a lower τ. Note that, in general, being able to predict rankings with an accuracy of $\tau = -1$ is as difficult as predicting rankings with an accuracy of $\tau = 1$, whereas a completely random ranking would have an expected value of $\tau = 0$. The range is therefore said to be symmetric.

However, there are two distinct ways of measuring rank correlation using Kendall's τ, related to the way *ties* are treated. They greatly affect how Kendall's τ numbers are to be interpreted, and especially the symmetry property. We explain the difference in detail in what follows.

Kendall's τ with ties penalised If the goal is to measure to what extent the difference in quality visible to a human annotator has been captured by an automatically produced hypothesis (recall-oriented view), then proposing a tie between t_1 and t_2 (t_1-equal-to-t_2) when the pair was judged (in the reference) as t_1-better-than-t_2 is treated as a failure-to-recall. In other words, it is as bad as proposing t_1-worse-than-t_2. Henceforth, we call this recall-oriented measure "Kendall's τ with ties penalised". This metric has the following properties:

- it is completely fair when comparing different methods to produce ranking hypotheses, because the denominator (number of total pairs) is the same (it is the number of non-tied pairs under the human judgements).
- it is non-symmetric, in the sense that a value of $\tau = -1$ is not as difficult to obtain as $\tau =$

[14] If the size $|S|$ is not divisible by n, then the last quantile S_n is assumed to contain the rest of the entries.

1 (simply proposing only ties gets a $\tau = -1$); hence, the sign of the τ value matters.

- the expected value of a completely random ranking is not necessarily $\tau = 0$, but rather depends on the number of ties in the reference rankings (i.e., it is test set dependent).

Kendall's τ with ties ignored If the goal is to measure to what extent the difference in quality signalled by an automatically produced hypothesis is reflected in the human annotation (precision-oriented view), then proposing t_1-equal-to-t_2 when the pair was judged differently in the reference does no harm the metric.

Henceforth, we call this precision-oriented measure "Kendall's τ with ties ignored". This metric has the following properties:

- it is not completely fair when comparing different methods to produce ranking hypotheses, because the denominator (number of total pairs) may not be the same (it is the number of non-tied pairs under each system's proposal).

- it is symmetric, in the sense that a value of $\tau = -1$ is as difficult to obtain as $\tau = 1$; hence, the sign of the τ value may not matter. [15]

- the expected value of a completely random ranking is $\tau = 0$ (test-set independent).

The first property is the most worrisome from the perspective of reporting the results of a shared task, because a system may fare very well on this metric simply because it choses not to commit (proposes ties) most of the time. Therefore, to give a better understanding of the systems' performance, for Kendall's τ with ties ignored we also provide the number of non-ties proposed by each system.

Task 1.3 Predicting post-editing time Submissions are evaluated in terms of Mean Average Error (MAE) against the actual time spent by post-editors (in seconds). By using a linear error measure we limit the influence of outliers: sentences that took very long to edit or where the measurement taken is questionable.

To further analyse the influence of extreme values, we also compute Spearman's rank correlation ρ coefficient which does not depend on the absolute values of the predictions.

We also give RMSE and Pearson's correlation coefficient r for reference.

Task 2 Predicting word-level scores The word-level task is primarily evaluated by macro-averaged F-measure. Because the class distribution is skewed – in the test data about one third of the tokens are marked as correct – we compute precision and recall and F_1 for each class individually. Consider the following confusion matrix for the two classes *Keep* and *Change*:

		predicted	
		(K)eep	(C)hange
expected	(K)eep	10	20
	(C)hange	30	40

For the given example we derive true-positive (tp), true-negative (tn), false-positive (fp), and false-negative (fn) counts:

$$\begin{array}{lll} tp_K = 10 & fp_K = 30 & fn_K = 20 \\ tp_C = 40 & fp_C = 20 & fn_C = 30 \end{array}$$

$$\text{precision}_K = \frac{tp_K}{tp_K + fp_K} = 10/40$$

$$\text{recall}_K = \frac{tp_K}{tp_K + fn_K} = 10/30$$

$$F_{1,K} = \frac{2 \cdot \text{precision}_K \cdot \text{recall}_K}{\text{precision}_K + \text{recall}_K}$$

A single cumulative statistic can be computed by averaging the resulting F-measures (*macro averaging*) or by *micro averaging* in which case precision and recall are first computed by accumulating the relevant values for all classes (Özgür et al., 2005), e.g.

$$\text{precision} = \frac{tp_K + tp_C}{(tp_K + fp_K) + (tp_C + fp_C)}$$

The latter gives equal weight to each example and is therefore dominated by performance on the largest class while macro-averaged F-measure gives equal weight to each class.

The same setup is used to evaluate the performance in the multiclass setting. Please note that here the test data only contains 4% examples for class (D)elete.

[15] In real life applications this distinction matters. Even if, from a computational perspective, it is as hard to get τ close to -1 as it is to get it close to 1, knowing the sign is the difference between selecting the best or the worse translation.

ID	Participating team
CMU	Carnegie Mellon University, USA (Hildebrand and Vogel, 2013)
CNGL	Centre for Next Generation Localization, Ireland (Bicici, 2013b)
DCU	Dublin City University, Ireland (Almaghout and Specia, 2013)
DCU-SYMC	Dublin City University & Symantec, Ireland (Rubino et al., 2013b)
DFKI	German Research Centre for Artificial Intelligence, Germany (Avramidis and Popovic, 2013)
FBK-UEdin	Fondazione Bruno Kessler, Italy & University of Edinburgh, UK (Camargo de Souza et al., 2013)
LIG	Laboratoire d'Informatique Grenoble, France (Luong et al., 2013)
LIMSI	Laboratoire d'Informatique pour la Mécanique et les Sciences de l'Ingénieur, France (Singh et al., 2013)
LORIA	Lorraine Laboratory of Research in Computer Science and its Applications, France (Langlois and Smaili, 2013)
SHEF	University of Sheffield, UK (Beck et al., 2013)
TCD-CNGL	Trinity College Dublin & CNGL, Ireland (Moreau and Rubino, 2013)
TCD-DCU-CNGL	Trinity College Dublin, Dublin City University & CNGL, Ireland (Moreau and Rubino, 2013)
UMAC	University of Macau, China (Han et al., 2013)
UPC	Universitat Politecnica de Catalunya, Spain (Formiga et al., 2013b)

Table 11: Participants in the WMT13 Quality Estimation shared task.

6.7 Participants

Table 11 lists all participating teams submitting systems to any subtask in this shared task. Each team was allowed up to two submissions for each subtask. In the descriptions below participation in specific tasks is denoted by a task identifier: T1.1, T1.2, T1.3, and T2.

Sentence-level baseline system (T1.1, T1.3): QuEst was used to extract 17 system-independent features from the source and translation files and the SMT training corpus that were found to be relevant in previous work (same features as in the WMT12 shared task):

- 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 probability of source and target sentences using language models provided by the task.
- average number of translations per source word in the sentence: as given by IBM 1 model thresholded so that $P(t|s) > 0.2$, and so that $P(t|s) > 0.01$ weighted by the inverse frequency of each word in the source side of the SMT training corpus.
- percentage of unigrams, bigrams and trigrams in frequency quartiles 1 (lower frequency words) and 4 (higher frequency words) in the source side of the SMT training corpus
- percentage of unigrams in the source sentence seen in the source side of the SMT training corpus.

These features are used to train a Support Vector Machine (SVM) regression algorithm using a radial basis function kernel within the SCIKIT-LEARN toolkit. The γ, ϵ and C parameters were optimized using a grid-search and 5-fold cross validation on the training set. We note that although the system is referred to as a "baseline", it is in fact a strong system. For tasks of the same type as 1.1 and 1.3, it has proved robust across a range of language pairs, MT systems, and text domains for predicting post-editing effort, as it has also been shown in the previous edition of the task (Callison-Burch et al., 2012).

The same features could be useful for a baseline system for Task 1.2. In our official results, however, the baseline for Task 1.2 is simpler than that: it proposes random ranks for each pair of alternative translations for a given source sentence, as we will discuss in Section 6.8.

CMU (T1.1, T1.2, T1.3): The CMU quality estimation system was trained on features based on language models, the MT system's distortion model and phrase table features, statistical word lexica, several sentence length statistics, source language word and bi-gram frequency statistics, n-best list agreement and diversity, source language parse, source-target word alignment and a dependency parse based cohesion penalty. These features were extracted using GIZA++, a forced alignment algorithm and the Stanford parser (de Marneffe et al., 2006). The prediction models were trained using four classifiers in the Weka toolkit (Hall et al., 2009): linear regression, M5P trees, multi layer perceptron and SVM regression. In addition to main system submission, a classic n-best list re-ranking approach was used for Task 1.2.

CNGL (T1.1, T1.2, T1.3, T2): CNGL systems are based on referential translation machines (RTM) (Biçici and van Genabith, 2013), parallel feature decay algorithms (FDA) (Bicici, 2013a), and machine translation performance predictor (MTPP) (Biçici et al., 2013), all of which allow to obtain language and MT system-independent predictions. For each task, RTM models were developed using the parallel corpora and the language model corpora distributed by the WMT13 translation task and the language model corpora provided by LDC for English and Spanish.

The sentence-level features are described in MTPP (Biçici et al., 2013); they include monolingual or bilingual features using n-grams defined over text or common cover link (CCL) (Seginer, 2007) structures as the basic units of information over which similarity calculations are made. RTMs use 308 features about coverage and diversity, IBM1, and sentence translation performance, retrieval closeness and minimum Bayes retrieval risk, distributional similarity and entropy, IBM2 alignment, character n-grams, and sentence readability. The learning models are Support Vector Machines (SVR) and SVR with partial least squares (SVRPLS).

The word-level features include CCL links, word length, location, prefix, suffix, form, context, and alignment, totalling 511K features for binary classification, and 637K for multiclass classification. Generalised linear models (GLM) (Collins, 2002) and GLM with dynamic learning (GLMd) were used.

DCU (T1.2): The main German-English submission uses six Combinatory Categorial Grammar (CCG) features: CCG supertag language model perplexity and log probability, the number of maximal CCG constituents in the translation output which are the highest-probability minimum number of CCG constituents that span the translation output, the percentage of CCG argument mismatches between each subsequent CCG supertags, the percentage of CCG argument mismatches between each subsequent CCG maximal categories and the minimum number of phrases detected in the translation output. A second submission uses the aforementioned CCG features combined with 80 features from QuEst as described in (Specia, 2011). For the CCG features, the C&C parser was used to parse the translation output. Moses was used to build the phrase table from the SMT training corpus with maximum phrase length set to 7. The language model of supertags was built using the SRILM toolkit. As learning algorithm, Logistic Regression as provided by the SCIKIT-LEARN toolkit was used. The training data was prepared by converting each ranking of translation outputs to a set of pairwise comparisons according to the approach proposed by Avramidis et al. (2011). The rankings were generated back from pairwise comparisons predicted by the model.

DCU-SYMC (T1.1): The DCU-Symantec team employed a wide set of features which included language model, n-gram counts and word-alignment features as well as syntactic features, topic model features and pseudo-reference features. The main learning algorithm was SVR, but regression tree learning was used to perform feature selection, reducing the initial set of 442 features to 96 features (DCU-Symantec alltypes) and 134

(DCU-Symantec combine). Two methods for feature selection were used: a best-first search in the feature space using regression trees to evaluate the subsets, and reading binarised features directly from the nodes of pruned regression trees.

The following NLP tools were used in feature extraction: the Brown English Wall-Street-Journal-trained statistical parser (Charniak and Johnson, 2005), a Lexical Functional Grammar parser (XLE), together with a hand-crafted Lexical Functional Grammar, the English ParGram grammar (Kaplan et al., 2004), and the TreeTagger part-of-speech tagger (Schmidt, 1994) with off-the-shelf publicly available pre-trained tagging models for English and Spanish. For pseudo-reference features, the Bing, Moses and Systran translation systems were used. The Mallet toolkit (McCallum, 2002) was used to build the topic models and features based on a grammar checker were extracted with LanguageTool.[16]

DFKI (T1.2, T1.3): DFKI's submission for Task 1.2 was based on decomposing rankings into pairs (Avramidis, 2012), where the best system for each pair was predicted with Logistic Regression (LogReg). For German-English, LogReg was trained with Stepwise Feature Selection (Hosmer, 1989) on two feature sets: *Feature Set 24* includes basic counts augmented with PCFG parsing features (number of VPs, alternative parses, parse probability) on both source and target sentences (Avramidis et al., 2011), and pseudo-reference METEOR score; the most successful set, *Feature Set 33* combines those 24 features with the 17 baseline features. For English-Spanish, LogReg was used with L2 Regularisation (Lin et al., 2007) and two feature sets were devised after scoring features with ReliefF (Kononenko, 1994) and Information Gain (Hunt et al., 1966). *Feature Set 431* combines 30 features with highest absolute Relief-F and Information Gain (15 from each). features with the highest

Task 1.3 was modelled using feature sets selected after Relief-F scoring of external black-box and glass-box features extracted from the SMT decoding process. The most successful submission (linear6) was trained with Linear Regression including the 17 features with highest positive Relief-F. Most prominent features include the alternative possible parses of the source and target sentence, the positions of the phrases with the lowest and highest probability and future cost estimate in the translation, the counts of phrases in the decoding graph whose probability or whether the future cost estimate is higher/lower than their standard deviation, counts of verbs and determiners, etc. The second submission (pls8) was trained with Partial Least Squares regression (Stone and Brooks, 1990) including more glass-box features.

FBK-Uedin (T1.1, T1.3):

The submissions explored features built on MT engine resources including automatic word alignment, n-best candidate translation lists, back-translations and word posterior probabilities. Information about word alignments is used to extract quantitative (amount and distribution of the alignments) and qualitative (importance of the aligned terms) features under the assumption that alignment information can help tasks where sentence-level semantic relations need to be identified (Souza et al., 2013). Three similar English-Spanish systems are built and used to provide pseudo-references (Soricut et al., 2012) and back-translations, from which automatic MT evaluation metrics could be computed and used as features.

All features were computed over a concatenation of several publicly available parallel corpora for the English-Spanish language pair such as Europarl, News Commentary, and MultiUN. The models were developed using supervised learning algorithms: SVMs (with feature selection step prior to model learning) and extremely randomized trees.

LIG (T2): The LIG systems are designed to deal with both binary and multiclass variants of the word level task. They integrate several features including: system-based (graph topology, language model, alignment context, etc.), lexical (Part-of-Speech tags), syntactic (constituent label, distance to the con-

[16] http://www.languagetool.org/

stituent tree root) and semantic (target and source polysemy count). Besides the existing components of the SMT system, feature extraction requires further external tools and resources, such as: TreeTagger (for POS tagging), Bekerley Parser trained with AnCora treebank (for generating constituent trees in Spanish), WordNet and BabelNet (for polysemy count), Google Translate. The feature set is then combined and trained using a Conditional Random Fields (CRF) learning method. During the labelling phase, the optimal threshold is tuned using a small development set split from the original training set. In order to retain the most informative features and eliminate the redundant ones, a Sequential Backward Selection algorithm is employed over the all-feature systems. With the binary classifier, the Boosting technique is applied to allow a number of sub feature sets to complement each other, resulting in the "stronger" combined system.

LIMSI (T1.1, T1.3): The two tasks were treated as regression problems using a simple elastic regression, a linear model trained with L_1 and L_2 regularisers. Regarding features, the submissions mainly aimed at evaluating the usefulness for quality estimation of n-gram posterior probabilities (Gispert et al., 2013) that quantify the probability for a given n-gram to be part of the system output. Their computation relies on all the hypotheses considered by a SMT system during decoding: intuitively, the more hypotheses a n-gram appears in, the more confident the system is that this n-gram is part of the correct translation, and the higher its posterior probability is. The feature set contains 395 other features that differs, in two ways, from the traditional features used in quality estimation. First, it includes several features based on large span continuous space language models (Le et al., 2011) that have already proved their efficiency both for the translation task and the quality estimation task. Second, each feature was expanded into two "normalized forms" in which their value was divided either by the source length or the target length and, when relevant, into a "ratio form" in which the feature value computed on the target sentence is divided by its value computed in the source sentence.

LORIA (T1.1): The system uses the 17 baseline features, plus several numerical and boolean features computed from the source and target sentences (Langlois et al., 2012). These are based on language model information (perplexity, level of back-off, intra-lingual triggers), translation table (IBM1 table, inter-lingual triggers). For language models, forward and backward models are built. Each feature gives a score to each word in the sentence, and the score of the sentence is the average of word scores. For several features, the score of a word depends on the score of its neighbours. This leads to 66 features. Support Vector Machines are used to learn a regression model. In training is done in a multi-stage procedure aimed at increasing the size of the training corpus. Initially, the training corpus with machine translated sentences provided by the task is used to train an SVM model. Then this model is applied to the post-edited and reference sentences (also provided as part of the task). These are added to the quality estimation training corpus using as labels the SVM predictions. An algorithm to tune the predicted scores on a development corpus is used.

SHEF (T1.1, T1.3): These submissions use Gaussian Processes, a non-parametric probabilistic learning framework for regression, along with two techniques to improve prediction performance and minimise the amount of resources needed for the problem: feature selection based on optimised hyperparameters and active learning to reduce the training set size (and therefore the annotation effort). The initial set features contains all black box and glass box features available within the QuEst framework (Specia et al., 2013) for the dataset at hand (160 in total for Task 1.1, and 80 for Task 1.3). The query selection strategy for active learning is based on the informativeness of the instances using Information Density, a measure that leverages between the variance among instances and how dense the region (in the feature space) where the instance is located is. To perform feature selection, following (Shah et al., 2013) features are ranked by the Gaussian Process algorithm according to their learned length scales, which can be interpreted as the relevance of such feature for the model. This information was used for feature selection by discarding the lowest ranked (least useful) ones. based on empirical results found in (Shah et al., 2013), the top 25 features for both models were selected and used to retrain the same regression algorithm.

UPC (T1.2): The methodology used a broad set of features, mainly available through the last version of the *Asiya* toolkit for MT evaluation (Gonzàlez et al., 2012)[17]. Concretely, 86 features were derived for the German-to-English and 97 features for the English-to-Spanish tasks. These features cover different approaches and include standard quality estimation features, as provided by the above mentioned *Asiya* and QUEST toolkits, but also a variety of features based on *pseudo-references*, explicit semantic analysis and specialised language models trained on the parallel and monolingual corpora provided by the WMT Translation Task.

The system selection task is approached by means of pairwise ranking decisions. It uses Random Forest classifiers with ties, expanding the work of 402013cFormiga et al.), from which a full ranking can be derived and the best system per sentence is identified. Once the classes are given by the Random Forest, one can build a graph by means of the adjacency matrix of the pairwise decision. The final ranking is assigned through a dominance scheme similar to Pighin et al. (2012).

An important remark of the methodology is the feature selection process, since it was noticed that the learner was sensitive to the features used. Selecting the appropriate set of features was crucial to achieve a good performance. The best feature combination was composed of: *i*) a baseline quality estimation feature set (Asiya or Quest) but not both of them, *ii*) Length Model, *iii*) Pseudo-reference aligned based features, and *iv*) adapted language models. However, within the *de-en* task, substituting Length Model and Aligned Pseudo-references by the features based on Semantic Roles could bring marginally better accuracy.

TCD-CNGL (T1.1) and **TCD-DCU-CNGL** (T1.3): The system is based on features which are commonly used for style classification (e.g. author identification). The assumption is that low/high quality translations can be characterised by some patterns which are frequent and/or differ significantly from the opposite category. Such features are intended to focus on striking patterns rather than to capture the global quality in a sentence, but they are used in conjunction with classical features for quality estimation (language modelling, etc.). This requires two steps in the training process: first the reference categories against which sentences will be compared are built, then the standard quality estimation model training stage is performed. Both datasets (Tasks 1.1 and 1.3) were used for both tasks. Since the number of features can be very high (up to 65,000), a combination of various heuristics for selecting features was used before the training stage (the submitted systems were trained using SVM with RBF kernels).

UMAC (T1.1, T1.2, T2): For Task 1.1, the feature set consists in POS sequences of the source and target languages, using 12 universal tags that are common in both languages. The algorithm is an enhanced version of the BLEU metric (EBLEU) designed with a modified length penalty and added recall factor, and having the precision and recall components grouped using the harmonic mean. For Task 1.2, in addition to the universal POS sequences of the source and target languages, features include the scores of length penalty, precision, recall and rank. Variants of EBLEU with different strategies for alignment are used, as well as a Naïve Bayes classification algorithm. For Task 2, the features used are unigrams (from previous 4th to following 3rd tokens), bigrams (from previous 2nd to following 2nd tokens), skip bigrams (previous and next token), trigrams (from previous 2nd to following 2nd tokens). The learning algorithms are Conditional Random Fields and Naïve Bayes.

[17] `http://asiya.lsi.upc.edu/`

6.8 Results

In what follows we give the official results for all tasks followed by a discussion that highlights the main findings for each of the tasks.

Task 1.1 Predicting post-editing distance

Table 12 summarises the results for the **ranking variant** of the task. They are sorted from best to worse using the DeltaAvg metric scores as primary key and the Spearman's rank correlation scores as secondary key.

The winning submissions for the ranking variant of Task 1.1 are CNGL SVRPLS, with a DeltaAvg score of 11.09, and DCU-SYMC alltypes, with a DeltaAvg score of 10.13. While the former holds the higher score, the difference is not significant at the $p \leq 0.05$ level as estimated by a bootstrap resampling test.

Both submissions are better than the baseline system by a very wide margin, a larger relative improvement than that obtained in the corresponding WMT12 task. In addition, five submissions (out of 12 systems) scored significantly higher than the baseline system (systems above the middle gray area), which is a larger proportion than that in last year's task (only 3 out of 16 systems), indicating that this shared task succeeded in pushing the state-of-the-art performance to new levels.

In addition to the performance of the official submission, we report results obtained by two oracle methods: the gold-label HTER metric computed against the post-edited translations as reference (Oracle HTER), and the BLEU metric (1-BLEU to obtain the same range as HTER) computed against the same post-edited translations as reference (Oracle HBLEU). The "Oracle HTER" DeltaAvg score of 16.38 gives an upperbound in terms of DeltaAvg for the test set used in this evaluation. It indicates that, for this set, the difference in post-editing effort between the top quality quantiles and the overall quality is 16.38 on average. The oracle based on HBLEU gives a lower DeltaAvg score, which is expected since HTER was our actual gold label. However, it is still significantly higher than the score of the winning submission, which shows that there is significant room for improvement even by the highest scoring submissions.

The results for the **scoring variant** of the task are presented in Table 13, sorted from best to worse by using the MAE metric scores as primary key and the RMSE metric scores as secondary key.

According to MAE scores, the winning submission is SHEF FS (MAE = 12.42), which uses feature selection and a novel learning algorithm for the task, Gaussian Processes. The baseline system is measured to have an MAE of 14.81, with six other submissions having performances that are not different from the baseline at a statistically significant level, as shown by the gray area in the middle of Table 13). Nine submissions (out of 16) scored significantly higher than the baseline system (systems above the middle gray area), a considerably higher proportion of submissions as compared to last year (5 out of 19), which indicates that this shared task also succeeded in pushing the state-of-the-art performance to new levels in terms of absolute scoring. Only one (6%) system scored significantly lower than the baseline, as opposed to 8 (42%) in last year's task.

For the sake of completeness, we also show oracles figures using the same methods as for the ranking variant of the task. Here the lowerbound in error (Oracle HTER) will clearly be zero, as both MAE and RMSE are measured against the same gold label used for the oracle computation. "Oracle HBLEU" is also not indicative in this case, as the although the values for the two metrics (HTER and HBLEU) are within the same ranges, they are not directly comparable. This explains the larger MAE/RMSE figures for "Oracle HBLEU" than those for most submissions.

Task 1.2 Selecting the best translation

Below we present the results for this task for each of the two Kendall's τ flavours presented in Section 6.6, for the German-English test set (Tables 14 and 16) and the English-Spanish test set (Tables 15 and 17). The results are sorted from best to worse using each of the Kendall's τ metric flavours.

For German-English, the winning submission is DFKI's logRegFss33 entry, for both Kendall's τ with ties penalised and ties ignored, with $\tau = 0.31$ (since this submission has no ties, the two metrics give the same τ value). A trivial baseline that proposes random ranks (with ties allowed) has a Kendall's τ with ties penalised of -0.12 (as this metric penalises the system's ties that were non-ties in the reference), and a Kendall's τ with ties ignored of 0.08. Most of the submissions performed better than this simple baseline. More interestingly perhaps is the comparison between the best submission and the performance by an ora-

System ID	DeltaAvg	Spearman ρ
● CNGL SVRPLS	11.09	0.55
● DCU-SYMC alltypes	10.13	0.59
SHEF FS	9.76	0.57
CNGL SVR	9.88	0.51
DCU-SYMC combine	9.84	0.59
CMU noB	8.98	0.57
SHEF FS-AL	8.85	0.50
Baseline bb17 SVR	8.52	0.46
CMU full	8.23	0.54
LIMSI	8.15	0.44
TCD-CNGL open	6.03	0.33
TCD-CNGL restricted	5.85	0.31
UMAC	2.74	0.11
Oracle HTER	16.38	1.00
Oracle HBLEU	15.74	0.93

Table 12: Official results for the ranking variant of the WMT13 Quality Estimation Task 1.1. The winning submissions are indicated by a ● (they are significantly better than all other submissions according to bootstrap resampling (10k 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. Oracle results that use human-references are also shown for comparison purposes.

System ID	MAE	RMSE
● SHEF FS	12.42	15.74
SHEF FS-AL	13.02	17.03
CNGL SVRPLS	13.26	16.82
LIMSI	13.32	17.22
DCU-SYMC combine	13.45	16.64
DCU-SYMC alltypes	13.51	17.14
CMU noB	13.84	17.46
CNGL SVR	13.85	17.28
FBK-UEdin extra	14.38	17.68
FBK-UEdin rand-svr	14.50	17.73
LORIA inctrain	14.79	18.34
Baseline bb17 SVR	14.81	18.22
TCD-CNGL open	14.81	19.00
LORIA inctraincont	14.83	18.17
TCD-CNGL restricted	15.20	19.59
CMU full	15.25	18.97
UMAC	16.97	21.94
Oracle HTER	0.00	0.00
Oracle HBLEU (1-HBLEU)	16.85	19.72

Table 13: Official results for the scoring variant of the WMT13 Quality Estimation Task 1.1. The winning submission is indicated by a ● (it is significantly better than the other submissions according to bootstrap resampling (10k 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. Oracle results that use human-references are also shown for comparison purposes.

German-English System ID	Kendall's τ with ties penalised
• DFKI logRegFss33	0.31
DFKI logRegFss24	0.28
CNGL SVRPLSF1	0.17
CNGL SVRF1	0.17
DCU CCG	0.15
UPC AQE+SEM+LM	0.11
UPC AQE+LeM+ALGPR+LM	0.10
DCU baseline+CCG	0.00
Baseline Random-ranks-with-ties	-0.12
UMAC EBLEU-I	-0.39
UMAC NB-LPR	-0.49
Oracle Human	1.00
Oracle BLEU (margin 0.00)	0.19
Oracle BLEU (margin 0.01)	0.05
Oracle METEOR-ex (margin 0.00)	0.23
Oracle METEOR-ex (margin 0.01)	0.06

Table 14: Official results for the Task 1.2 of the WMT13 Quality Estimation shared task for German-English, using as metric Kendall's τ with ties penalised. The winning submissions are indicated by a •. Oracle results that use human-references are also shown for comparison purposes.

English-Spanish System ID	Kendall's τ with ties penalised
• CNGL SVRPLSF1	0.15
CNGL SVRF1	0.13
DFKI logRegL2-411	0.09
DFKI logRegL2-431	0.04
UPC QQE+LeM+ALGPR+LM	-0.03
UPC AQE+LeM+ALGPR+LM	-0.06
CMU BLEUopt	-0.11
Baseline Random-ranks-with-ties	-0.23
UMAC EBLEU-A	-0.27
UMAC EBLEU-I	-0.35
CMU cls	-0.63
Oracle Human	1.00
Oracle BLEU (margin 0.00)	0.17
Oracle BLEU (margin 0.02)	-0.06
Oracle METEOR-ex (margin 0.00)	0.19
Oracle METEOR-ex (margin 0.02)	0.05

Table 15: Official results for the Task 1.2 of the WMT13 Quality Estimation shared task for English-Spanish, using as metric Kendall's τ with ties penalised. The winning submissions are indicated by a •. Oracle results that use human-references are also shown for comparison purposes.

German-English System ID	Kendall's τ with ties ignored	Nr. of non-ties / Nr. of decisions
● DFKI logRegFss33	0.31	882/882
DFKI logRegFss24	0.28	882/882
UPC AQE+SEM+LM	0.27	768/882
UPC AQE+LeM+ALGPR+LM	0.24	788/882
DCU CCG	0.18	862/882
CNGL SVRPLSF1	0.17	882/882
CNGL SVRF1	0.17	881/882
Baseline Random-ranks-with-ties	0.08	718/882
DCU baseline+CCG	0.01	874/882
UMAC NB-LPR	0.01	447/882
UMAC EBLEU-I	-0.03	558/882
Oracle Human	1.00	882/882
Oracle BLEU (margin 0.00)	0.22	859/882
Oracle BLEU (margin 0.01)	0.27	728/882
Oracle METEOR-ex (margin 0.00)	0.20	869/882
Oracle METEOR-ex (margin 0.01)	0.24	757/882

Table 16: Official results for the Task 1.2 of the WMT13 Quality Estimation shared task for German-English, using as metric Kendall's τ with ties ignored. The winning submissions are indicated by a ●. Oracle results that use human-references are also shown for comparison purposes.

English-Spanish System ID	Kendall's τ with ties ignored	Nr. of non-ties / Nr. of decisions
● CMU cls	0.23	192/633
CNGL SVRPLSF1	0.16	632/633
CNGL SVRF1	0.13	631/633
DFKI logRegL2-411	0.13	610/633
UPC QQE+LeM+ALGPR+LM	0.11	554/633
UPC AQE+LeM+ALGPR+LM	0.08	554/633
UMAC EBLEU-A	0.07	430/633
DFKI logRegL2-431	0.04	633/633
Baseline Random-ranks-with-ties	0.03	507/633
UMAC EBLEU-I	0.02	407/633
CMU BLEUopt	-0.11	633/633
Oracle Human	1.00	633/633
Oracle BLEU (margin 0.00)	0.19	621/633
Oracle BLEU (margin 0.02)	0.26	474/633
Oracle METEOR-ex (margin 0.00)	0.25	623/633
Oracle METEOR-ex (margin 0.02)	0.28	517/633

Table 17: Official results for the Task 1.2 of the WMT13 Quality Estimation shared task for English-Spanish, using as metric Kendall's τ with ties ignored. The winning submissions are indicated by a ●. Oracle results that use human-references are also shown for comparison purposes.

cle method that has access to human-created references. This oracle uses human references to compute BLEU and METEOR scores for each translation segment, and consequently computes rankings for the competing translations based on these scores. To reflect the impact of ties on the two versions of Kendall's τ metric we use, we allow these ranks to be tied if the difference between the oracle BLEU or METEOR scores is smaller than a margin (see lower section of Tables 14 and 16, with margins of 0 and 0.01 for the scores). For example, under a regime of BLEU with margin 0.01, a translation with BLEU score of 0.172 would get the same rank as a translation with BLEU score of 0.164 (difference of 0.008), but a higher rank than a translation with BLEU score of 0.158 (difference of 0.014). Not surprisingly, under the Kendall's τ with ties penalised the best Oracle BLEU or METEOR performance happens for a 0.0 margin (which makes ties possible only for exactly-matching scores), for a value of $\tau = 0.19$ and $\tau = 0.23$, respectively. Under the Kendall's τ with ties ignored, the Oracle BLEU performance for a 0.01 margin (i.e, translations under 1 BLEU point should be considered as having the same rank) achieves $\tau = 0.27$, while Oracle METEOR for a 0.01 margin achieves $\tau = 0.24$. These values are lower than the $\tau = 0.31$ of the winning submission without access to reference translations, suggesting that quality estimation models are capable of better modelling translation differences compared to traditional, human reference-based MT evaluation metrics.

For English-Spanish, under Kendall's τ with ties penalised the winning submission is CNGL's SVRPLSF1, with $\tau = 0.15$. Under Kendall's τ with ties ignored, the best scoring submission is CMU's cls with $\tau = 0.23$, but this is achieved by offering non-tie judgements only for 192 of the 633 total judgements (30% of them). As we discussed in Section 6.6, the "Kendall's τ with ties ignored" metric is weak with respect to comparing different submissions, since it favours systems that are do not commit to a given rank and rather produce a large number of ties. This becomes even clearer when we look at the performance of the oracle methods (Tables 15 and 17). Under Kendall's τ with ties penalised, "Oracle BLEU" (margin 0.00) achieves $\tau = 0.17$, while under Kendall's τ with ties ignored, "Oracle BLEU" (margin 0.02) has a $\tau = 0.26$. This results in 474 non-tie decisions (75% of them), and a better τ value compared to "Oracle BLEU" (margin 0.00), with a $\tau = 0.19$ under the same metric. The oracle values for both BLEU and METEOR are close to the τ values of the winning submissions, supporting the conclusion that quality estimation techniques can successfully replace traditional, human reference-based MT evaluation metrics.

Task 1.3 Predicting post-editing time

Results for this task are presented in Table 18. A third of the submissions was able to beat the baseline. Among these FBK-UEDIN's submission ranked best in terms of MAE, our main metric for this task, and also achieved the lowest RMSE.

Only three systems were able to beat our baseline in terms of MAE. Please note that while all features were available to the participants, our baseline is actually a competitive system.

The second-best entry, CNGL SVR, reached the highest Spearman's rank correlation, our secondary metric. Furthermore, in terms of this metric all four top-ranking entries, two by CNGL and FBK-UEDIN respectively, are significantly better than the baseline (10k bootstrap resampling test with 95% confidence intervals). As high ranking submissions also yield strong rank correlation to the observed post-editing time, we can be confident that improvements in MAE are not only due to better handling of extreme cases.

Many participants submitted two variants of their systems with different numbers of features and/or machine learning approaches. In Table 18 we can see these are grouped closely together giving rise to the assumption that the general pool of available features and thereby the used resources and strongest features are most relevant for a system's performance. Another hint in that direction is the observation the top-ranked systems rely on additional data and resources to generate their features.

Task 2 Predicting word-level scores

Results for this task are presented in Table 19 and 20, sorted by macro average F_1. Since this is a new task, we have yet to establish a strong baseline. For reference we provide a trivial baseline that predicts the dominant class – *(K)eep* – for every token.

The first observation in Table 19 is that this trivial baseline is difficult to beat in terms of accuracy. However, considering our main metric – macro-

System ID	MAE	RMSE	Pearson's r	Spearman's ρ
● FBK-UEDIN Extra	47.5	82.6	0.65	0.75
● FBK-UEDIN Rand-SVR	47.9	86.7	0.66	0.74
CNGL SVR	49.2	90.4	0.67	0.76
CNGL SVRPLS	49.6	86.6	0.68	0.74
CMU slim	51.6	84.7	0.63	0.68
Baseline bb17 SVR	51.9	93.4	0.61	0.70
DFKI linear6	52.4	84.3	0.64	0.68
CMU full	53.6	92.2	0.58	0.60
DFKI pls8	53.6	88.3	0.59	0.67
TCD-DCU-CNGL SVM2	55.8	98.9	0.47	0.60
TCD-DCU-CNGL SVM1	55.9	99.4	0.48	0.60
SHEF FS	55.9	103.1	0.42	0.61
SHEF FS-AL	64.6	99.1	0.57	0.60
LIMSI elastic	70.6	114.4	0.58	0.64

Table 18: Official results for the Task 1.3 of the WMT13 Quality Estimation shared-task. The winning submissions are indicated by a ● (they are significantly better than all other submissions according to bootstrap resampling (10k 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.

			Keep			Change		
System ID	Accuracy	Prec.	Recall	F_1	Prec.	Recall	F_1	Macro F_1
● LIG FS_BIN	0.74	0.79	0.86	0.82	0.56	0.43	0.48	0.65
● LIG BOOST_BIN	0.74	0.78	0.88	0.83	0.57	0.37	0.45	0.64
CNGL GLM	0.70	0.76	0.86	0.80	0.47	0.31	0.38	0.59
UMAC NB	0.56	0.82	0.49	0.62	0.37	0.73	0.49	0.55
CNGL GLMd	0.71	0.74	0.93	0.82	0.51	0.19	0.28	0.55
UMAC CRF	0.71	0.72	0.98	0.83	0.49	0.04	0.07	0.45
Baseline (one class)	0.71	0.71	1.00	0.83	0.00	0.00	0.00	0.42

Table 19: Official results for Task 2: binary classification on word level of the WMT13 Quality Estimation shared-task. The winning submissions are indicated by a ●.

System ID	F_1 Keep	F_1 Substitute	F_1 Delete	Micro-F_1	Macro-F_1
● LIG FS_MULT	0.83	0.44	0.072	0.72	0.45
● LIG ALL_MULT	0.83	0.45	0.064	0.72	0.45
UMAC NB	0.62	0.43	0.042	0.52	0.36
CNGL GLM	0.83	0.18	0.028	0.71	0.35
CNGL GLMd	0.83	0.14	0.034	0.72	0.34
UMAC CRF	0.83	0.04	0.012	0.71	0.29
Baseline (one class)	0.83	0.00	0.000	0.71	0.28

Table 20: Official results for Task 2: multiclass classification on word level of the WMT13 Quality Estimation shared-task. The winning submissions are indicated by a ●.

average F_1 – it is clear that all systems outperform the baseline. The winning systems by LIG for the binary task are also the top ranking systems on the multiclass task.

While promising results are found for the binary variant of the task where systems are able to achieve an F_1 of almost 0.5 for the relevant class – *Change*, the multiclass prediction variant of the task seem to suffer from its severe class imbalance. In fact, none of the systems shows good performance when predicting deletions.

6.9 Discussion

In what follows, we discuss the main accomplishments of this shared task starting from the goals we had previously identified for it.

Explore various granularity levels for the quality-prediction task The decision on which level of granularity quality estimation is applied depends strongly on the intended application. In Task 2 we tested binary word-level classification in a post-editing setting. If such annotation is presented through a user interface we imagine that words marked as incorrect would be hidden from the editor, highlighted as possibly wrong or that a list of alternatives would we generated.

With respect to the poor improvements over trivial baselines, we consider that the results for word-level prediction could be mostly connected to limitations of the datasets provided, which are very small for word-level prediction, as compared to successful previous work such as (Bach et al., 2011). Despite the limited amount of training data, several systems were able to predict dubious words (binary variant of the task), showing that this can be a promising task. Extending the granularity even further by predicting the actual editing action necessary for a word yielded less positive results than the binary setting.

We cannot directly compare sentence- and word-level results. However, since sentence-level predictions can benefit from more information available and therefore more signal on which the prediction is based, the natural conclusion is that, if there is a choice in the prediction granularity, to opt for the coarser one possible (i.e., sentence-level over word-level). But certain applications may require finer granularity levels, and therefore word-level predictions can still be very valuable.

Explore the prediction of more objective scores Given the multitude of possible applications for quality estimation we must decide which predicted values are both useful and accurate. In this year's task we have attempted to address the usefulness criterion by moving from the subjective, human judgement-based scores, to the prediction of scores that can be more easily interpreted for practical applications: post-editing distance or types of edits (word-level), post-editing time, and ranking of alternative translations.

The general promise of using objective scores is that predicting a value that is related to the use case will make quality estimation more applicable and yield lower deviance compared to the use of proxy metrics. The magnitude of this benefit should be sufficient to account for the possible additional effort related to collecting such scores.

While a direct comparison between the different types of scores used for this year's tasks is not possible as they are based on different datasets, if we compare last year's task on predicting 1-5 likert scores (and generating an overall ranking of all translations in the test set) with this year's Task 1.1, which is virtually the same, but using post-editing distance as gold-label, we see that the number of systems that outperform the baseline [18] is proportionally larger this year. We can also notice a higher relative improvement of these submissions over the baseline system. While this could simply be a consequence of progress in the field, it may also provide an indication that objective metrics are more suitable for the problem.

Particularly with respect to post-editing time, given that this label has a long tailed distribution and is not trivial to measure even in a controlled environment, the results of Task 1.3 are encouraging. Comparison with the better results seen on Tasks 1.1 and 1.2, however, suggests that, for Task 1.3, additional data processing, filtering, and modelling (including modelling translator-specific traits such as their variance in time) is required, as evidenced in (Cohn and Specia, 2013).

Explore the use of quality estimation techniques to replace reference-based MT evaluation metrics When it comes to the task of automatically ranking alternative translations generated by different MT systems, the traditional use of reference-based MT evaluation metrics is challenged by the findings of this task.

The top ranking quality estimation submissions

[18] The two baselines are exactly the same, and therefore the comparison is meaningful.

to Task 1.2 have performances that outperform or are at least at the same level with the ones that involve the use of human references. The most interesting property of these techniques is that, being reference-free, they can be used for any source sentences, and therefore are ready to be deployed for arbitrary texts.

An immediate application for this capability is a procedure by which MT system-selection is performed, based on the output of such quality estimators. Additional measurements are needed to determine the level of improvement in translation quality that the current performance of these techniques can achieve in a system-selection scenario.

Identify new and effective quality indicators Quality indicators, or features, are core to the problem of quality estimation. One significant difference this year with respect to previous year was the availability of QuEst, a framework for the extraction of a large number of features. A few submissions used these larger sets – as opposed to the 17 baseline features used in the 2012 edition – as their starting point, to which they added other features. Most features available in this framework, however, had already been used in previous work.

Novel families of features used this year which seems to have played an important role are those proposed by CNGL. They include a number of language and MT-system independent monolingual and bilingual similarity metrics between the sentences for prediction and corpora of the language pair under consideration. Based on standard regression algorithm (the same used by the baseline system), the submissions from CNGL using such feature families topped many of the tasks.

Another interesting family of features is that used by TCD-CNGL and TCD-DCU-CNGL for Tasks 1.1 and 1.3. These were borrowed from work on style or authorship identification. The assumption is that low/high quality translations can be characterised by some patterns which are frequent and/or differ significantly from patterns belonging to the opposite category.

Like in last year's task, the vast majority of the participating systems used external resources in addition to those provided for the task, particularly for linguistically-oriented features, such as parsers, part-of-speech taggers, named entity recognizers, etc. A novel set of syntactic features based on Combinatory Categorial Grammar (CCG) performed reasonably well in Task 1.2: with six CCG-based features and no additional features, the system outperformed the baseline system and also a second submission where the 17 baseline features were added. This highlights the potential of linguistically-motivated features for the problem.

As expected, different feature sets were used for different tasks. This is essential for Task 2, where word-level features are certainly necessary. For example, LIG used a number of lexical features such as part-of-speech tag, word-posterior probabilities, syntactic (constituent label, distance to the constituent tree root, and target and source polysemy count). For submissions where a sequence labelling algorithm such as a Conditional Random Fields was used for prediction, the interdependencies between adjacent words and labels was also modelled though features.

Pseudo-references, i.e., scores from standard evaluation metrics such as BLEU based on translations generated by an alternative MT system as "reference", featured in more than half of the submissions for sentence-level tasks. This is not surprising given their performance in previous work on quality estimation.

Identify effective machine learning techniques for all variants of the quality estimation task For the sentence-level tasks, standard regression methods such as SVR performed well as in the previous edition of the shared task, topping the results for the ranking variant of Task 1.1, both first and second place. In fact this algorithm was used by most submissions that outperformed the baseline. An alternative algorithm to SVR with very promising results and which was introduced for the problem this year is that of Gaussian Processes. It was used by SHEF, the winning submission in the scoring variant of Task 1.1, which also performed well in the ranking variant, despite its hyperparameters having been optimised for scoring only. Algorithms behave similarly for Task 1.3, with SVR performing particularly well.

For Task 1.2, logistic regression performed the best or among the best, along with SVR. One of the most effective approach for this task, however, appears to be one that is better tailored for the task, namely pair-wise decomposition for ranking. This approach benefits from transforming a k-way ranking problem into a series of simpler, 2-way ranking problems, which can be more accurately solved. Another approach that shows promise is that of ensemble of regressors, in which the output is the results combining the predictions of different regression models.

Linear-chain Conditional Random Fields are a popular model of choice for sequence labelling tasks and have been successfully used by several participants in Task 2, along with discriminatively trained Hidden Markov Models and Naïve Bayes.

As in the previous edition, feature engineering and feature selection prior to model learning were important components in many submissions. However, the role of individual features is hard to judge separately from the role of the machine learning techniques employed.

Establish the state of the art performance All four tasks addressed in this shared task have achieved a dual role that is important for the research community: (i) to make publicly available new data sets that can serve to compare different approaches and contributions; and (ii) to establish the present state-of-the-art performance in the field, so that progress can be easily measured and tracked. In addition, the public availability of the scoring scripts makes evaluation and direct comparison straightforward.

Many participants submitted predictions for several tasks. Comparison of the results shows that there is little overlap between the best systems when the predicted value is varied. While we did not formally require the participants to use similar systems across tasks, these results indicate that specialised systems with features selected depending on the predicted variable can in fact be beneficial.

As we mentioned before, compared to the previous edition of the task, we noticed (for Task 1.1) a larger relative improvement of scores over the baseline system, as well as a larger proportion of systems outperforming the baseline systems, which are a good indication that the field is progressing over the years. For example, in the scoring variant of Task 1.1, last year only 5 out of 20 systems (i.e. 25% of the systems) were able to significantly outperform the baseline. This year, 9 out 16 systems (i.e. 56%) outperformed the same baseline. Last year, the relative improvement of the winning submission with respect to the baseline system was 13%, while this year the relative improvement is of 19%.

Overall, the tables of results presented in Section 6.8 give a comprehensive view of the current state-of-the-art on the data sets used for this shared task, as well as indications on how much room there still is for improvement via figures from oracle methods. As a result, people interested in contributing to research in these machine translation quality estimation tasks will be able to do so in a principled way, with clearly established state-of-the-art levels and straightforward means of comparison.

7 Summary

As in previous incarnations of this workshop we carried out an extensive manual and automatic evaluation of machine translation performance, and we used the human judgements that we collected to validate automatic metrics of translation quality. We also refined last year's quality estimation task, asking for methods that predict sentence-level post-editing effort and time, rank translations from alternative systems, and pinpoint words in the output that are more likely to be wrong.

As in previous years, all data sets generated by this workshop, including the human judgments, system translations and automatic scores, are publicly available for other researchers to analyze.[19]

Acknowledgments

This work was supported in parts by the MosesCore, Casmacat, Khresmoi, Matecat and QTLaunchPad projects funded by the European Commission (7th Framework Programme), and by gifts from Google, Microsoft and Yandex.

We would also like to thank our colleagues Matouš Macháček and Martin Popel for detailed discussions.

References

Allauzen, A., Pécheux, N., Do, Q. K., Dinarelli, M., Lavergne, T., Max, A., Le, H.-S., and Yvon, F. (2013). LIMSI @ WMT13. In *Proceedings of the Eighth Workshop on Statistical Machine Translation*, pages 60–67, Sofia, Bulgaria. Association for Computational Linguistics.

Almaghout, H. and Specia, L. (2013). A CCG-based Quality Estimation Metric for Statistical Machine Translation. In *Proceedings of MT Summit XIV (to appear)*, Nice, France.

Ammar, W., Chahuneau, V., Denkowski, M., Hanneman, G., Ling, W., Matthews, A., Murray,

[19] `http://statmt.org/wmt13/results.html`

K., Segall, N., Lavie, A., and Dyer, C. (2013). The CMU machine translation systems at WMT 2013: Syntax, synthetic translation options, and pseudo-references. In *Proceedings of the Eighth Workshop on Statistical Machine Translation*, pages 68–75, Sofia, Bulgaria. Association for Computational Linguistics.

Avramidis, E. (2012). Comparative Quality Estimation: Automatic Sentence-Level Ranking of Multiple Machine Translation Outputs. In *Proceedings of 24th International Conference on Computational Linguistics*, pages 115–132, Mumbai, India.

Avramidis, E. and Popovic, M. (2013). Selecting feature sets for comparative and time-oriented quality estimation of machine translation output. In *Proceedings of the Eighth Workshop on Statistical Machine Translation*, pages 327–334, Sofia, Bulgaria. Association for Computational Linguistics.

Avramidis, E., Popović, M., Vilar, D., and Burchardt, A. (2011). Evaluate with confidence estimation: Machine ranking of translation outputs using grammatical features. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*.

Aziz, W., Mitkov, R., and Specia, L. (2013). Ranking Machine Translation Systems via Post-Editing. In *Proc. of Text, Speech and Dialogue (TSD)*, Lecture Notes in Artificial Intelligence, Berlin / Heidelberg. Západočeská univerzita v Plzni, Springer Verlag.

Bach, N., Huang, F., and Al-Onaizan, Y. (2011). Goodness: A method for measuring machine translation confidence. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 211–219, Portland, Oregon, USA.

Beck, D., Shah, K., Cohn, T., and Specia, L. (2013). 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.

Biçici, E., Groves, D., and van Genabith, J. (2013). Predicting sentence translation quality using extrinsic and language independent features. *Machine Translation*.

Biçici, E. and van Genabith, J. (2013). CNGL-CORE: Referential translation machines for measuring semantic similarity. In **SEM 2013: The Second Joint Conference on Lexical and Computational Semantics*, Atlanta, Georgia, USA. Association for Computational Linguistics.

Bicici, E. (2013a). Feature decay algorithms for fast deployment of accurate statistical machine translation systems. In *Proceedings of the Eighth Workshop on Statistical Machine Translation*, pages 76–82, Sofia, Bulgaria. Association for Computational Linguistics.

Bicici, E. (2013b). Referential translation machines for quality estimation. In *Proceedings of the Eighth Workshop on Statistical Machine Translation*, pages 341–349, Sofia, Bulgaria. Association for Computational Linguistics.

Bílek, K. and Zeman, D. (2013). CUni multilingual matrix in the WMT 2013 shared task. In *Proceedings of the Eighth Workshop on Statistical Machine Translation*, pages 83–89, Sofia, Bulgaria. Association for Computational Linguistics.

Bojar, O., Kos, K., and Mareček, D. (2010). Tackling Sparse Data Issue in Machine Translation Evaluation. In *Proceedings of the ACL 2010 Conference Short Papers*, pages 86–91, Uppsala, Sweden. Association for Computational Linguistics.

Bojar, O., Rosa, R., and Tamchyna, A. (2013). Chimera – three heads for English-to-Czech translation. In *Proceedings of the Eighth Workshop on Statistical Machine Translation*, pages 90–96, Sofia, Bulgaria. Association for Computational Linguistics.

Borisov, A., Dlougach, J., and Galinskaya, I. (2013). Yandex school of data analysis machine translation systems for WMT13. In *Proceedings of the Eighth Workshop on Statistical Machine Translation*, pages 97–101, Sofia, Bulgaria. 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 metaevaluation 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., Buck, C., Turchi, M., and Negri, M. (2013). FBK-UEdin participation to the WMT13 quality estimation shared task. In *Proceedings of the Eighth Workshop on Statistical Machine Translation*, pages 350–356, Sofia, Bulgaria. Association for Computational Linguistics.

Charniak, E. and Johnson, M. (2005). Coarse-to-fine n-best parsing and maxent discriminative reranking. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, pages 173–180. Association for Computational Linguistics.

Cho, E., Ha, T.-L., Mediani, M., Niehues, J., Herrmann, T., Slawik, I., and Waibel, A. (2013). The Karlsruhe Institute of Technology translation systems for the WMT 2013. In *Proceedings of the Eighth Workshop on Statistical Machine Translation*, pages 102–106, Sofia, Bulgaria. Association for Computational Linguistics.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurment*, 20(1):37–46.

Cohn, T. and Specia, L. (2013). Modelling Annotator Bias with Multi-task Gaussian Processes: An Application to Machine Translation Quality Estimation. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (to appear)*.

Collins, M. (2002). Discriminative training methods for hidden markov models: theory and experiments with perceptron algorithms. In *Proceedings of the ACL-02 conference on Empirical methods in natural language processing - Volume 10*, EMNLP '02, pages 1–8, Stroudsburg, PA, USA. Association for Computational Linguistics.

de Marneffe, M.-C., MacCartney, B., and Manning, C. D. (2006). Generating typed dependency parses from phrase structure parses. In *Proceedings of LREC-06*, pages 449–454.

Denkowski, M. and Lavie, A. (2010). Meteor-next and the meteor paraphrase tables: improved evaluation support for five target languages. In *Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR*, WMT '10, pages 339–342, Stroudsburg, PA, USA. Association for Computational Linguistics.

Durgar El-Kahlout, I. and Mermer, C. (2013). TÜbtak-blgem german-english machine translation systems for w13. In *Proceedings of the Eighth Workshop on Statistical Machine Translation*, pages 107–111, Sofia, Bulgaria. Association for Computational Linguistics.

Durrani, N., Fraser, A., Schmid, H., Sajjad, H., and Farkas, R. (2013a). Munich-Edinburgh-Stuttgart submissions of OSM systems at WMT13. In *Proceedings of the Eighth Workshop on Statistical Machine Translation*, pages 120–125, Sofia, Bulgaria. Association for Computational Linguistics.

Durrani, N., Haddow, B., Heafield, K., and Koehn, P. (2013b). Edinburgh's machine translation systems for European language pairs. In *Proceedings of the Eighth Workshop on Statistical Machine Translation*, pages 112–119, Sofia, Bulgaria. Association for Computational Linguistics.

Eidelman, V., Wu, K., Ture, F., Resnik, P., and Lin, J. (2013). Towards efficient large-scale feature-rich statistical machine translation. In *Proceedings of the Eighth Workshop on Statistical Machine Translation*, pages 126–131, Sofia, Bulgaria. Association for Computational Linguistics.

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.

Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5):378–382.

Formiga, L., Costa-jussà, M. R., Mariño, J. B., Fonollosa, J. A. R., Barrón-Cedeño, A., and Marquez, L. (2013a). The TALP-UPC phrase-based translation systems for WMT13: System combination with morphology generation, domain adaptation and corpus filtering. In *Proceedings of the Eighth Workshop on Statistical Machine Translation*, pages 132–138, Sofia, Bulgaria. Association for Computational Linguistics.

Formiga, L., Gonzàlez, M., Barrón-Cedeño, A., Fonollosa, J. A. R., and Marquez, L. (2013b). The TALP-UPC approach to system selection: Asiya features and pairwise classification using random forests. In *Proceedings of the Eighth Workshop on Statistical Machine Translation*, pages 357–362, Sofia, Bulgaria. Association for Computational Linguistics.

Formiga, L., Màrquez, L., and Pujantell, J. (2013c). Real-life translation quality estimation for mt system selection. In *Proceedings of MT Summit XIV (to appear)*, Nice, France.

Galuščáková, P., Popel, M., and Bojar, O. (2013). PhraseFix: Statistical post-editing of TectoMT. In *Proceedings of the Eighth Workshop on Statistical Machine Translation*, pages 139–145, Sofia, Bulgaria. Association for Computational Linguistics.

Gispert, A., Blackwood, G., Iglesias, G., and Byrne, W. (2013). N-gram posterior probability confidence measures for statistical machine translation: an empirical study. *Machine Translation*, 27:85–114.

Gonzàlez, M., Giménez, J., and Màrquez, L. (2012). A graphical interface for mt evaluation and error analysis. In *Proceedings of the ACL 2012 System Demonstrations*, pages 139–144, Jeju Island, Korea.

Green, S., Cer, D., Reschke, K., Voigt, R., Bauer, J., Wang, S., Silveira, N., Neidert, J., and Manning, C. D. (2013). Feature-rich phrase-based translation: Stanford University's submission to the WMT 2013 translation task. In *Proceedings of the Eighth Workshop on Statistical Machine Translation*, pages 146–151, Sofia, Bulgaria. Association for Computational Linguistics.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., and Witten, I. H. (2009). The weka data mining software: an update. *SIGKDD Explor. Newsl.*, 11(1):10–18.

Han, A. L.-F., Wong, D. F., Chao, L. S., Lu, Y., He, L., Wang, Y., and Zhou, J. (2013). A description of tunable machine translation evaluation systems in WMT13 metrics task. In *Proceedings of the Eighth Workshop on Statistical Machine Translation*, pages 412–419, Sofia, Bulgaria. Association for Computational Linguistics.

Hildebrand, S. and Vogel, S. (2013). MT quality estimation: The CMU system for WMT'13. In *Proceedings of the Eighth Workshop on Statistical Machine Translation*, pages 371–377, Sofia, Bulgaria. Association for Computational Linguistics.

Hosmer, D. (1989). *Applied logistic regression*. Wiley, New York, 8th edition.

Huet, S., Manishina, E., and Lefèvre, F. (2013). Factored machine translation systems for Russian-English. In *Proceedings of the Eighth Workshop on Statistical Machine Translation*, pages 152–155, Sofia, Bulgaria. Association for Computational Linguistics.

Hunt, E., Martin, J., and Stone, P. (1966). *Experiments in Induction*. Academic Press, New York.

Kaplan, R., Riezler, S., King, T., Maxwell, J., Vasserman, A., and Crouch, R. (2004). Speed and accuracy in shallow and deep stochastic parsing. In *Proceedings of the Human Language Technology Conference and the 4th Annual Meeting of the North American Chapter of the Association for Computational Linguistics (HLT/NAACL 04)*.

Kauchak, D. and Barzilay, R. (2006). Paraphrasing for automatic evaluation. In *Proceedings of the main conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics*, HLT-NAACL '06, pages 455–462, Stroudsburg, PA, USA. Association for Computational Linguistics.

Koehn, P. (2012). Simulating human judgment in machine translation evaluation campaigns. In *International Workshop on Spoken Language Translation (IWSLT)*.

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.

Kononenko, I. (1994). Estimating attributes: analysis and extensions of RELIEF. In *Proceedings of the European conference on machine learning on Machine Learning*, pages 171–182, Secaucus, NJ, USA. Springer-Verlag New York, Inc.

Kos, K. and Bojar, O. (2009). Evaluation of Machine Translation Metrics for Czech as the Target Language. *Prague Bulletin of Mathematical Linguistics*, 92:135–147.

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

Langlois, D., Raybaud, S., and Smaïli, K. (2012). Loria system for the wmt12 quality estimation shared task. In *Proceedings of the Seventh Workshop on Statistical Machine Translation*, pages 114–119, Montréal, Canada.

Langlois, D. and Smaili, K. (2013). LORIA system for the WMT13 quality estimation shared task. In *Proceedings of the Eighth Workshop on Statistical Machine Translation*, pages 378–383, Sofia, Bulgaria. Association for Computational Linguistics.

Le, H. S., Oparin, I., Allauzen, A., Gauvain, J.-L., and Yvon, F. (2011). Structured output layer neural network language model. In *ICASSP*, pages 5524–5527.

Lin, C.-J., Weng, R. C., and Keerthi, S. S. (2007). Trust region Newton methods for large-scale logistic regression. In *Proceedings of the 24th international conference on Machine learning - ICML '07*, pages 561–568, New York, New York, USA. ACM Press.

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.

Macháček, M. and Bojar, O. (2013). Results of the WMT13 metrics shared task. In *Proceedings of the Eighth Workshop on Statistical Machine Translation*, pages 43–49, Sofia, Bulgaria. Association for Computational Linguistics.

Matusov, E. and Leusch, G. (2013). Omnifluent English-to-French and Russian-to-English systems for the 2013 Workshop on Statistical Machine Translation. In *Proceedings of the Eighth Workshop on Statistical Machine Translation*, pages 156–161, Sofia, Bulgaria. Association for Computational Linguistics.

McCallum, A. K. (2002). MALLET: a machine learning for language toolkit.

Miceli Barone, A. V. and Attardi, G. (2013). Pre-reordering for machine translation using transition-based walks on dependency parse trees. In *Proceedings of the Eighth Workshop on Statistical Machine Translation*, pages 162–167, Sofia, Bulgaria. Association for Computational Linguistics.

Moreau, E. and Rubino, R. (2013). An approach using style classification features for quality estimation. In *Proceedings of the Eighth Workshop on Statistical Machine Translation*, pages 427–432, Sofia, Bulgaria. Association for Computational Linguistics.

Nadejde, M., Williams, P., and Koehn, P. (2013). Edinburgh's syntax-based machine translation systems. In *Proceedings of the Eighth Workshop on Statistical Machine Translation*, pages 168–174, Sofia, Bulgaria. Association for Computational Linguistics.

Okita, T., Liu, Q., and van Genabith, J. (2013). Shallow semantically-informed PBSMT and HPBSMT. In *Proceedings of the Eighth Workshop on Statistical Machine Translation*, pages 175–182, Sofia, Bulgaria. Association for Computational Linguistics.

Özgür, A., Özgür, L., and Güngör, T. (2005). Text categorization with class-based and corpus-based keyword selection. In *Proceedings of the 20th International Conference on Computer and Information Sciences*, ISCIS'05, pages 606–615, Berlin, Heidelberg. Springer.

Peitz, S., Mansour, S., Huck, M., Freitag, M., Ney, H., Cho, E., Herrmann, T., Mediani, M., Niehues, J., Waibel, A., Allauzen, A., Khanh Do, Q., Buschbeck, B., and Wandmacher, T. (2013a). Joint WMT 2013 submission of the QUAERO project. In *Proceedings of the Eighth Workshop on Statistical Machine Translation*, pages 183–190, Sofia, Bulgaria. Association for Computational Linguistics.

Peitz, S., Mansour, S., Peter, J.-T., Schmidt, C., Wuebker, J., Huck, M., Freitag, M., and Ney, H. (2013b). The RWTH aachen machine translation system for WMT 2013. In *Proceedings of the Eighth Workshop on Statistical Machine Translation*, pages 191–197, Sofia, Bulgaria. Association for Computational Linguistics.

Pighin, D., Formiga, L., and Màrquez, L. (2012). A graph-based strategy to streamline translation quality assessments. In *Proceedings of the Tenth Conference of the Association for Machine Translation in the Americas (AMTA'2012)*, San Diego, USA.

Pino, J., Waite, A., Xiao, T., de Gispert, A., Flego, F., and Byrne, W. (2013). The University of Cambridge Russian-English system at WMT13. In *Proceedings of the Eighth Workshop on Statistical Machine Translation*, pages 198–203, Sofia, Bulgaria. Association for Computational Linguistics.

Post, M., Ganitkevitch, J., Orland, L., Weese, J., Cao, Y., and Callison-Burch, C. (2013). Joshua 5.0: Sparser, better, faster, server. In *Proceedings of the Eighth Workshop on Statistical Machine Translation*, pages 204–210, Sofia, Bulgaria. Association for Computational Linguistics.

Rubino, R., Toral, A., Cortés Vaíllo, S., Xie, J., Wu, X., Doherty, S., and Liu, Q. (2013a). The CNGL-DCU-Prompsit translation systems for WMT13. In *Proceedings of the Eighth Workshop on Statistical Machine Translation*, pages 211–216, Sofia, Bulgaria. Association for Computational Linguistics.

Rubino, R., Wagner, J., Foster, J., Roturier, J., Samad Zadeh Kaljahi, R., and Hollowood, F. (2013b). DCU-Symantec at the WMT 2013 quality estimation shared task. In *Proceedings of the Eighth Workshop on Statistical Machine Translation*, pages 390–395, Sofia, Bulgaria. Association for Computational Linguistics.

Sajjad, H., Smekalova, S., Durrani, N., Fraser, A., and Schmid, H. (2013). QCRI-MES submission at WMT13: Using transliteration mining to improve statistical machine translation. In *Proceedings of the Eighth Workshop on Statistical Machine Translation*, pages 217–222, Sofia, Bulgaria. Association for Computational Linguistics.

Schmidt, H. (1994). Probabilistic part-of-speech tagging using decision trees. In *Proceedings of the International Conference on New Methods in Natural Language Processing*.

Seginer, Y. (2007). *Learning Syntactic Structure*. PhD thesis, University of Amsterdam.

Shah, K., Cohn, T., and Specia, L. (2013). An Investigation on the Effectiveness of Features for Translation Quality Estimation. In *Proceedings of MT Summit XIV (to appear)*, Nice, France.

Singh, A. K., Wisniewski, G., and Yvon, F. (2013). LIMSI submission for the WMT'13 quality estimation task: an experiment with n-gram posteriors. In *Proceedings of the Eighth Workshop on Statistical Machine Translation*, pages 396–402, Sofia, Bulgaria. Association for Computational Linguistics.

Smith, J., Saint-Amand, H., Plamada, M., Koehn, P., Callison-Burch, C., and Lopez, A. (2013). Dirt cheap web-scale parallel text from the Common Crawl. In *Proceedings of the 2013 Conference of the Association for Computational Linguistics (ACL 2013)*, Sofia, Bulgaria. Association for Computational Linguistics.

Snover, M., Dorr, B., Schwartz, R., Micciulla, L., and Makhoul, J. (2006). 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.

Snover, M., Madnani, N., Dorr, B. J., and Schwartz, R. (2009). Fluency, adequacy, or hter?: exploring different human judgments with a tunable mt metric. In *Proceedings of the Fourth Workshop on Statistical Machine Trans-*

lation, StatMT '09, pages 259–268, Stroudsburg, PA, USA. Association for Computational Linguistics.

Soricut, R., Bach, N., and Wang, Z. (2012). The SDL Language Weaver Systems in the WMT12 Quality Estimation Shared Task. In *Proceedings of the 7th Workshop on Statistical Machine Translation*, pages 145–151.

Souza, J. G. C. d., Espl-Gomis, M., Turchi, M., and Negri, M. (2013). Exploiting qualitative information from automatic word alignment for cross-lingual nlp tasks. In *The 51st Annual Meeting of the Association for Computational Linguistics - Short Papers (ACL Short Papers 2013)*.

Specia, L. (2011). Exploiting Objective Annotations for Measuring Translation Post-editing Effort. In *Proceedings of the 15th Conference of the European Association for Machine Translation*, pages 73–80, Leuven.

Specia, L., Shah, K., de Souza, J. G. C., and Cohn, T. (2013). QuEst - A Translation Quality Estimation Framework. In *Proceedings of the 51th Conference of the Association for Computational Linguistics (ACL), Demo Session*, Sofia, Bulgaria. Association for Computational Linguistics.

Stolcke, A. (2002). SRILM – An extensible language modeling toolkit. In *Proceedings of the 7th International Conference on Spoken Language Processing (ICSLP 2002)*, pages 901–904.

Stone, M. and Brooks, R. J. (1990). Continuum regression: cross-validated sequentially constructed prediction embracing ordinary least squares, partial least squares and principal components regression. *Journal of the Royal Statistical Society Series B Methodological*, 52(2):237–269.

Stymne, S., Hardmeier, C., Tiedemann, J., and Nivre, J. (2013). Tunable distortion limits and corpus cleaning for SMT. In *Proceedings of the Eighth Workshop on Statistical Machine Translation*, pages 223–229, Sofia, Bulgaria. Association for Computational Linguistics.

Tantug, A. C., Oflazer, K., and El-Kahlout, I. D. (2008). BLEU+: a Tool for Fine-Grained BLEU Computation. In (ELRA), E. L. R. A., editor, *Proceedings of the Sixth International Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco.

Weller, M., Kisselew, M., Smekalova, S., Fraser, A., Schmid, H., Durrani, N., Sajjad, H., and Farkas, R. (2013). Munich-Edinburgh-Stuttgart submissions at WMT13: Morphological and syntactic processing for SMT. In *Proceedings of the Eighth Workshop on Statistical Machine Translation*, pages 230–237, Sofia, Bulgaria. Association for Computational Linguistics.

A Pairwise System Comparisons by Human Judges

Tables 21–30 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.

	UEDIN-HEAFIELD	ONLINE-B	MES	UEDIN	ONLINE-A	UEDIN-SYNTAX	CU-ZEMAN	CU-TAMCHYNA	DCU-FDA	JHU	SHEF-WPROA
UEDIN-HEAFIELD	–	.50	.48†	.43‡	.47†	.43‡	.44‡	.38‡	.32‡	.25‡	.26‡
ONLINE-B	.50	–	.46‡	.48†	.47†	.49	.44‡	.40‡	.39‡	.29‡	.27‡
MES	**.52†**	**.54‡**	–	.49	.47⋆	.44‡	.45‡	.42‡	.41‡	.27‡	.25‡
UEDIN	**.57‡**	**.52†**	**.51**	–	**.51**	.48†	.47‡	.42‡	.39‡	.28‡	.25‡
ONLINE-A	**.53†**	**.53†**	**.53⋆**	.49	–	.48	**.51**	.44‡	.42‡	.31‡	.30‡
UEDIN-SYNTAX	**.57‡**	**.51**	**.56‡**	**.52†**	**.52**	–	**.51**	.43‡	.41‡	.29‡	.26‡
CU-ZEMAN	**.56‡**	**.56‡**	**.55‡**	**.53‡**	.49	.49	–	.45‡	.42‡	.32‡	.29‡
CU-TAMCHYNA	**.62‡**	**.60‡**	**.58‡**	**.58‡**	**.56‡**	**.57‡**	**.55‡**	–	.46‡	.35‡	.32‡
DCU-FDA	**.68‡**	**.61‡**	**.59‡**	**.61‡**	**.58‡**	**.59‡**	**.58‡**	**.54‡**	–	.32‡	.32‡
JHU	**.75‡**	**.71‡**	**.73‡**	**.72‡**	**.69‡**	**.71‡**	**.68‡**	**.65‡**	**.68‡**	–	.46‡
SHEF-WPROA	**.74‡**	**.73‡**	**.75‡**	**.75‡**	**.70‡**	**.74‡**	**.71‡**	**.68‡**	**.68‡**	**.54‡**	–
score	.60	.58	.57	.56	.54	.54	.53	.48	.45	.32	.29
rank	1	2-3	2-4	3-5	4-7	5-7	6-7	8	9	10	11

Table 21: Head to head comparison, ignoring ties, for Czech-English systems

	CU-BOJAR	CU-DEPFIX	ONLINE-B	UEDIN	CU-ZEMAN	MES	ONLINE-A	CU-PHRASEFIX	CU-TECTOMT	COMMERCIAL-1	COMMERCIAL-2	SHEF-WPROA
CU-BOJAR	–	**.51**	.47†	.44‡	.42‡	.43‡	.48	.41‡	.37‡	.39‡	.38‡	.33‡
CU-DEPFIX	.49	–	.48⋆	.42‡	.43‡	.41‡	.47†	.42‡	.40‡	.40‡	.39‡	.34‡
ONLINE-B	**.53†**	**.52⋆**	–	.47‡	.44‡	.44‡	.44‡	.44‡	.44‡	.41‡	.36‡	.34‡
UEDIN	**.56‡**	**.58‡**	**.53‡**	–	.47†	.47‡	.48	.45‡	.44‡	.42‡	.43‡	.38‡
CU-ZEMAN	**.58‡**	**.57‡**	**.56‡**	**.53†**	–	.49	.49	.48†	.46‡	.47‡	.47‡	.35‡
MES	**.57‡**	**.59‡**	**.56‡**	**.53‡**	**.51**	–	.50	.47†	.46‡	.43‡	.44‡	.42‡
ONLINE-A	**.52**	**.53†**	**.56‡**	**.52**	**.51**	.50	–	**.52**	.47⋆	.47†	.47†	.46†
CU-PHRASEFIX	**.59‡**	**.58‡**	**.56‡**	**.55‡**	**.52†**	**.53†**	.48	–	.49	.48†	.49	.42‡
CU-TECTOMT	**.63‡**	**.60‡**	**.56‡**	**.56‡**	**.54‡**	**.54‡**	**.53⋆**	**.51**	–	.46‡	.46‡	.40‡
COMMERCIAL-1	**.61‡**	**.60‡**	**.59‡**	**.58‡**	**.53‡**	**.57‡**	**.53†**	**.52†**	**.54‡**	–	.49	.42‡
COMMERCIAL-2	**.62‡**	**.61‡**	**.64‡**	**.57‡**	**.53‡**	**.56‡**	**.53†**	**.51**	**.54‡**	**.51**	–	.43‡
SHEF-WPROA	**.67‡**	**.66‡**	**.66‡**	**.62‡**	**.65‡**	**.58‡**	**.54†**	**.58‡**	**.60‡**	**.58‡**	**.57‡**	–
score	.58	.57	.56	.52	.50	.50	.49	.48	.47	.45	.45	.38
rank	1-2	1-2	3	4	5-7	5-7	5-8	7-9	8-9	10-11	10-11	12

Table 22: Head to head comparison, ignoring ties, for English-Czech systems

	ONLINE-B	ONLINE-A	UEDIN-SYNTAX	UEDIN	QUAERO	KIT	MES	RWTH-JANE	MES-SZEGED-REORDER-SPLIT	LIMSI-NCODE-SOUL	TUBITAK	UMD	DCU	CU-ZEMAN	JHU	SHEF-WPROA	DESRT
ONLINE-B	–	.48	.44‡	.37‡	.44‡	.41‡	.42‡	.40‡	.35‡	.37‡	.32‡	.31‡	.31‡	.27‡	.23‡	.18‡	.16‡
ONLINE-A	**.52**	–	.47	.45†	.47	.43‡	.42‡	.41‡	.44‡	.40‡	.35‡	.36‡	.34‡	.31‡	.27‡	.25‡	.21‡
UEDIN-SYNTAX	**.56‡**	**.53**	–	.48	.46†	.48⋆	.46†	.46†	.45‡	.45‡	.35‡	.35‡	.34‡	.28‡	.25‡	.20‡	.19‡
UEDIN	**.63‡**	**.55†**	**.52**	–	**.51**	.46‡	.47†	.49	.44‡	.43‡	.39‡	.34‡	.35‡	.32‡	.28‡	.24‡	.22‡
QUAERO	**.56‡**	**.53**	**.54†**	.49	–	.49	**.52**	.44‡	.46‡	.44‡	.39‡	.38‡	.37‡	.30‡	.31‡	.25‡	.21‡
KIT	**.59‡**	**.57‡**	**.52⋆**	**.54‡**	**.51**	–	.45‡	**.51**	.43‡	.46‡	.37‡	.38‡	.41‡	.35‡	.31‡	.25‡	.21‡
MES	**.58‡**	**.58‡**	**.54†**	**.53†**	.48	**.55‡**	–	.49	.49	.46‡	.44‡	.37‡	.40‡	.34‡	.30‡	.26‡	.20‡
RWTH-JANE	**.60‡**	**.59‡**	**.54†**	**.51**	**.56‡**	.49	**.51**	–	.46‡	.50	.45‡	.46‡	.47†	.38‡	.33‡	.28‡	.20‡
MES-SZEGED-REORDER-SPLIT	**.65‡**	**.56‡**	**.55‡**	**.56‡**	**.54‡**	**.57‡**	**.51**	**.54‡**	–	**.53⋆**	.44‡	.41‡	.41‡	.36‡	.34‡	.31‡	.21‡
LIMSI-NCODE-SOUL	**.63‡**	**.60‡**	**.55‡**	**.57‡**	**.56‡**	**.54‡**	**.54‡**	.50	.47⋆	–	**.51**	.45‡	.43‡	.37‡	.34‡	.30‡	.22‡
TUBITAK	**.68‡**	**.65‡**	**.65‡**	**.61‡**	**.61‡**	**.63‡**	**.56‡**	**.55‡**	**.56‡**	.49	–	.48⋆	.49	.39‡	.41‡	.30‡	.25‡
UMD	**.69‡**	**.64‡**	**.65‡**	**.66‡**	**.62‡**	**.62‡**	**.63‡**	**.54‡**	**.59‡**	**.55‡**	**.52⋆**	–	.48⋆	.41‡	.40‡	.33‡	.27‡
DCU	**.69‡**	**.66‡**	**.66‡**	**.65‡**	**.63‡**	**.59‡**	**.60‡**	**.53†**	**.59‡**	**.57‡**	**.51**	**.52⋆**	–	.41‡	.38‡	.37‡	.25‡
CU-ZEMAN	**.73‡**	**.69‡**	**.72‡**	**.68‡**	**.70‡**	**.65‡**	**.66‡**	**.62‡**	**.64‡**	**.63‡**	**.61‡**	**.59‡**	**.59‡**	–	.44‡	.43‡	.29‡
JHU	**.77‡**	**.73‡**	**.75‡**	**.72‡**	**.69‡**	**.69‡**	**.70‡**	**.67‡**	**.66‡**	**.66‡**	**.59‡**	**.60‡**	**.62‡**	**.56‡**	–	.43‡	.30‡
SHEF-WPROA	**.82‡**	**.75‡**	**.80‡**	**.76‡**	**.75‡**	**.75‡**	**.74‡**	**.72‡**	**.69‡**	**.70‡**	**.70‡**	**.67‡**	**.63‡**	**.57‡**	**.57‡**	–	.41‡
DESRT	**.84‡**	**.79‡**	**.81‡**	**.78‡**	**.79‡**	**.79‡**	**.80‡**	**.80‡**	**.79‡**	**.78‡**	**.75‡**	**.73‡**	**.75‡**	**.71‡**	**.70‡**	**.59‡**	–
score	.66	.62	.60	.58	.58	.57	.56	.54	.53	.52	.48	.46	.46	.39	.36	.31	.23
rank	1	2-3	2-3	4-5	4-5	5-7	6-7	8-9	8-10	9-10	11	12-13	12-13	14	15	16	17

Table 23: Head to head comparison, ignoring ties, for German-English systems

	ONLINE-B	PROMT	UEDIN-SYNTAX	ONLINE-A	UEDIN	KIT	STANFORD	LIMSI-NCODE-SOUL	MES-REORDER	JHU	CU-ZEMAN	TUBITAK	UU	SHEF-WPROA	RWTH-JANE
ONLINE-B	–	**.55‡**	.50	.45⋆	.45‡	.34‡	.37‡	.37‡	.37‡	.32‡	.32‡	.33‡	.24‡	.21‡	.26‡
PROMT	.45‡	–	.48⋆	.50	.43‡	.40‡	.39‡	.36‡	.37‡	.31‡	.31‡	.32‡	.27‡	.24‡	.27‡
UEDIN-SYNTAX	.50	**.52⋆**	–	**.57†**	.45‡	.43‡	.38‡	.41‡	.39‡	.38‡	.33‡	.33‡	.26‡	.25‡	.22‡
ONLINE-A	**.55⋆**	.50	.43†	–	**.51**	.42†	.48	.41‡	.36‡	.44⋆	.44⋆	.38‡	.32‡	.27‡	.29‡
UEDIN	**.55‡**	**.57‡**	**.55‡**	.49	–	**.52**	.45‡	.45‡	.42‡	.43‡	.37‡	.34‡	.29‡	.27‡	.31‡
KIT	**.66‡**	**.60‡**	**.57‡**	**.58†**	.48	–	.48	.45‡	.42‡	.36‡	.39‡	.40‡	.30‡	.29‡	.26‡
STANFORD	**.63‡**	**.61‡**	**.62‡**	**.52**	**.55‡**	**.52**	–	.50	.44‡	.48	.44‡	.43‡	.34‡	.29‡	.32‡
LIMSI-NCODE-SOUL	**.63‡**	**.64‡**	**.59‡**	**.59‡**	**.55‡**	**.55‡**	.50	–	.44‡	.44‡	.44‡	.47†	.40‡	.34‡	.33‡
MES-REORDER	**.63‡**	**.63‡**	**.61‡**	**.64‡**	**.58‡**	**.58‡**	**.56‡**	**.56‡**	–	.50	.46‡	.49	.38‡	.37‡	.34‡
JHU	**.68‡**	**.69‡**	**.62‡**	**.56⋆**	**.57‡**	**.64‡**	**.52**	**.56‡**	.50	–	.48⋆	.45‡	.36‡	.37‡	.34‡
CU-ZEMAN	**.68‡**	**.69‡**	**.67‡**	**.56⋆**	**.63‡**	**.61‡**	**.56‡**	**.56‡**	**.54‡**	**.52⋆**	–	.48	.40‡	.33‡	.34‡
TUBITAK	**.67‡**	**.68‡**	**.67‡**	**.62‡**	**.66‡**	**.60‡**	**.57‡**	**.53†**	**.51**	**.55‡**	**.52**	–	.38‡	.40‡	.32‡
UU	**.76‡**	**.73‡**	**.74‡**	**.68‡**	**.71‡**	**.70‡**	**.66‡**	**.60‡**	**.62‡**	**.64‡**	**.60‡**	**.62‡**	–	.44‡	.46†
SHEF-WPROA	**.79‡**	**.76‡**	**.75‡**	**.73‡**	**.73‡**	**.71‡**	**.71‡**	**.66‡**	**.63‡**	**.63‡**	**.67‡**	**.60‡**	**.56‡**	–	.47†
RWTH-JANE	**.74‡**	**.73‡**	**.78‡**	**.71‡**	**.69‡**	**.74‡**	**.68‡**	**.67‡**	**.66‡**	**.66‡**	**.66‡**	**.68‡**	**.54†**	**.53†**	–
score	.63	.63	.61	.58	.57	.55	.52	.50	.47	.47	.46	.45	.36	.32	.32
rank	1-2	1-2	3	3-5	4-6	5-6	7	8	9-11	9-11	10-12	11-12	13	14-15	14-15

Table 24: Head to head comparison, ignoring ties, for English-German systems

	UEDIN-HEAFIELD	UEDIN	ONLINE-B	LIMSI-NCODE-SOUL	KIT	ONLINE-A	MES-SIMPLIFIEDFRENCH	DCU	RWTH	CMU-TREE-TO-TREE	CU-ZEMAN	JHU	SHEF-WPROA
UEDIN-HEAFIELD	–	.45‡	.46‡	.46‡	.42‡	.42‡	.34‡	.34‡	.29‡	.33‡	.31‡	.28‡	.24‡
UEDIN	**.55‡**	–	**.52⋆**	.43‡	.45‡	.46⋆	.40‡	.38‡	.33‡	.36‡	.33‡	.32‡	.23‡
ONLINE-B	**.54‡**	.48⋆	–	.49	.46‡	.44‡	.45‡	.40‡	.38‡	.34‡	.36‡	.31‡	.26‡
LIMSI-NCODE-SOUL	**.54‡**	**.57‡**	**.51**	–	**.52⋆**	.47	.45‡	.42‡	.38‡	.36‡	.34‡	.31‡	.28‡
KIT	**.58‡**	**.55‡**	**.54‡**	.48⋆	–	.47	.46‡	.44‡	.39‡	.38‡	.37‡	.33‡	.28‡
ONLINE-A	**.58‡**	**.54⋆**	**.56‡**	**.53**	**.53**	–	.47	.45†	.40‡	.40‡	.39‡	.34‡	.32‡
MES-SIMPLIFIEDFRENCH	**.66‡**	**.60‡**	**.55‡**	**.55‡**	**.54‡**	**.53**	–	.48⋆	.44‡	.40‡	.39‡	.39‡	.32‡
DCU	**.66‡**	**.62‡**	**.60‡**	**.58‡**	**.56‡**	**.55†**	**.52⋆**	–	.45‡	.45‡	.42‡	.41‡	.36‡
RWTH	**.71‡**	**.67‡**	**.62‡**	**.62‡**	**.61‡**	**.60‡**	**.56‡**	**.55‡**	–	.48⋆	.47†	.47⋆	.38‡
CMU-TREE-TO-TREE	**.67‡**	**.64‡**	**.66‡**	**.64‡**	**.62‡**	**.60‡**	**.60‡**	**.55‡**	**.52⋆**	–	.50	.48	.37‡
CU-ZEMAN	**.69‡**	**.67‡**	**.64‡**	**.66‡**	**.63‡**	**.61‡**	**.61‡**	**.58‡**	**.53†**	.50	–	.47†	.39‡
JHU	**.72‡**	**.68‡**	**.69‡**	**.69‡**	**.67‡**	**.66‡**	**.61‡**	**.59‡**	**.53⋆**	**.52**	**.53†**	–	.45‡
SHEF-WPROA	**.76‡**	**.77‡**	**.74‡**	**.72‡**	**.72‡**	**.68‡**	**.68‡**	**.64‡**	**.62‡**	**.63‡**	**.61‡**	**.55‡**	–
score	.63	.60	.59	.57	.56	.54	.51	.48	.43	.42	.42	.38	.32
rank	1	2-3	2-3	4-5	4-5	5-6	7	8	9-10	9-11	10-11	12	13

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

	UEDIN	ONLINE-B	LIMSI-NCODE-SOUL	KIT	PROMT	STANFORD	MES	MES-INFLECTION	RWTH-PHRASE-BASED-JANE	ONLINE-A	DCU	CU-ZEMAN	JHU	OMNIFLUENT	ITS-LATL	ITS-LATL-PE
UEDIN	–	.49	.47⋆	.48	.50	.44‡	.41‡	.40‡	.47⋆	.39‡	.41‡	.35‡	.29‡	.30‡	.27‡	.24‡
ONLINE-B	**.51**	–	.46‡	.47⋆	.47†	.44‡	.49	.43‡	.43‡	.43‡	.38‡	.35‡	.36‡	.28‡	.25‡	.25‡
LIMSI-NCODE-SOUL	**.53⋆**	**.54‡**	–	.45‡	.48	.48	.45‡	.43‡	.44‡	.45†	.41‡	.32‡	.34‡	.30‡	.27‡	.27‡
KIT	**.52**	**.53⋆**	**.55‡**	–	.48	.46†	.45‡	.43‡	.45‡	.46⋆	.38‡	.30‡	.33‡	.31‡	.29‡	.29‡
PROMT	.50	**.53†**	**.52**	**.52**	–	.50	.48	**.52⋆**	.45‡	.47	.48⋆	.38‡	.36‡	.36‡	.34‡	.31‡
STANFORD	**.56‡**	**.56‡**	**.52**	**.54†**	.50	–	**.52**	.48	.44‡	.49	.44‡	.39‡	.34‡	.36‡	.30‡	.29‡
MES	**.59‡**	**.51**	**.55‡**	**.55‡**	**.52**	.48	–	**.52**	**.51**	.45⋆	.45‡	.36‡	.37‡	.34‡	.29‡	.29‡
MES-INFLECTION	**.60‡**	**.57‡**	**.57‡**	**.57‡**	.48⋆	**.52**	.48	–	**.54†**	**.51**	.46†	.37‡	.35‡	.31‡	.33‡	.31‡
RWTH-PHRASE-BASED-JANE	**.53⋆**	**.57‡**	**.56‡**	**.55‡**	**.55‡**	**.56‡**	.49	.46†	–	**.53**	.49	.38‡	.36‡	.34‡	.35‡	.31‡
ONLINE-A	**.61‡**	**.57‡**	**.55†**	**.54⋆**	**.53**	**.51**	**.55⋆**	.49	.47	–	.50	.45†	.38‡	.38‡	.39‡	.35‡
DCU	**.59‡**	**.62‡**	**.59‡**	**.62‡**	**.52⋆**	**.56‡**	**.55‡**	**.54†**	**.51**	.50	–	.42‡	.40‡	.40‡	.36‡	.35‡
CU-ZEMAN	**.65‡**	**.65‡**	**.68‡**	**.70‡**	**.62‡**	**.61‡**	**.64‡**	**.63‡**	**.62‡**	**.55†**	**.58‡**	–	.50	.42‡	.41‡	.37‡
JHU	**.71‡**	**.64‡**	**.66‡**	**.67‡**	**.64‡**	**.66‡**	**.63‡**	**.65‡**	**.64‡**	**.62‡**	**.60‡**	.50	–	.47‡	.42‡	.38‡
OMNIFLUENT	**.70‡**	**.72‡**	**.70‡**	**.69‡**	**.64‡**	**.64‡**	**.66‡**	**.69‡**	**.66‡**	**.62‡**	**.60‡**	**.58‡**	**.53‡**	–	.43‡	.42‡
ITS-LATL	**.73‡**	**.75‡**	**.72‡**	**.71‡**	**.66‡**	**.70‡**	**.71‡**	**.67‡**	**.65‡**	**.61‡**	**.64‡**	**.59‡**	**.58‡**	**.57‡**	–	.45‡
ITS-LATL-PE	**.76‡**	**.75‡**	**.73‡**	**.71‡**	**.69‡**	**.71‡**	**.71‡**	**.69‡**	**.69‡**	**.65‡**	**.65‡**	**.63‡**	**.62‡**	**.58‡**	**.55‡**	–
score	.60	.60	.58	.58	.55	.55	.54	.53	.53	.51	.49	.42	.40	.38	.35	.32
rank	1-2	1-3	2-4	3-4	5-7	5-8	5-8	6-9	7-10	9-11	10-11	12	13	14	15	16

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

	UEDIN-HEAFIELD	ONLINE-B	UEDIN	ONLINE-A	MES	LIMSI-NCODE-SOUL	DCU	DCU-OKITA	DCU-FDA	CU-ZEMAN	JHU	SHEF-WPROA
UEDIN-HEAFIELD	–	.49	.42‡	.45⋆	.43‡	.40‡	.34‡	.43‡	.37‡	.34‡	.31‡	.15‡
ONLINE-B	**.51**	–	.49	.44‡	.46‡	.47†	.42‡	.39‡	.40‡	.37‡	.37‡	.16‡
UEDIN	**.58‡**	**.51**	–	**.55†**	.50	.47†	.43‡	.42‡	.39‡	.39‡	.35‡	.14‡
ONLINE-A	**.55⋆**	**.56‡**	.45†	–	.50	.44‡	.45†	.42‡	.42‡	.41‡	.37‡	.18‡
MES	**.57‡**	**.54‡**	.50	.50	–	.47†	.45‡	.41‡	.41‡	.40‡	.38‡	.15‡
LIMSI-NCODE-SOUL	**.60‡**	**.53†**	**.53†**	**.56‡**	**.53†**	–	.46‡	.45‡	.44‡	.43‡	.38‡	.18‡
DCU	**.66‡**	**.58‡**	**.57‡**	**.55†**	**.55‡**	**.54‡**	–	.44‡	.47†	.42‡	.41‡	.16‡
DCU-OKITA	**.57‡**	**.61‡**	**.58‡**	**.58‡**	**.59‡**	**.55‡**	**.56‡**	–	.49	.46‡	.46‡	.18‡
DCU-FDA	**.63‡**	**.60‡**	**.61‡**	**.58‡**	**.59‡**	**.56‡**	**.53†**	**.51**	–	.48⋆	.43‡	.18‡
CU-ZEMAN	**.66‡**	**.63‡**	**.61‡**	**.59‡**	**.60‡**	**.57‡**	**.58‡**	**.54‡**	**.52⋆**	–	.43‡	.18‡
JHU	**.69‡**	**.63‡**	**.65‡**	**.63‡**	**.62‡**	**.62‡**	**.59‡**	**.54‡**	**.57‡**	**.57‡**	–	.22‡
SHEF-WPROA	**.85‡**	**.84‡**	**.86‡**	**.82‡**	**.85‡**	**.82‡**	**.84‡**	**.82‡**	**.82‡**	**.82‡**	**.78‡**	–
score	.62	.59	.57	.57	.56	.53	.51	.48	.48	.46	.42	.16
rank	1	2	3-5	3-5	3-5	6	7	8-9	8-9	10	11	12

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

	ONLINE-B	ONLINE-A	UEDIN	PROMT	MES	TALP-UPC	LIMSI-NCODE	DCU	DCU-FDA	DCU-OKITA	CU-ZEMAN	JHU	SHEF-WPROA
ONLINE-B	–	.49	.45‡	.43‡	.38‡	.35‡	.34‡	.35‡	.37‡	.34‡	.33‡	.32‡	.23‡
ONLINE-A	**.51**	–	.49	.48	.38‡	.46⋆	.42‡	.41‡	.43‡	.38‡	.38‡	.37‡	.31‡
UEDIN	**.55‡**	**.51**	–	.49	.46†	.45‡	.43‡	.42‡	.36‡	.38‡	.38‡	.38‡	.26‡
PROMT	**.57‡**	**.52**	**.51**	–	.46‡	.48	.43‡	.43‡	.40‡	.37‡	.39‡	.34‡	.29‡
MES	**.62‡**	**.62‡**	**.54†**	**.54‡**	–	.46‡	.44‡	.44‡	.41‡	.40‡	.43‡	.36‡	.32‡
TALP-UPC	**.65‡**	**.54⋆**	**.55‡**	**.52**	**.54‡**	–	.50	.45‡	.44‡	.40‡	.40‡	.37‡	.32‡
LIMSI-NCODE	**.66‡**	**.58‡**	**.57‡**	**.57‡**	**.56‡**	.50	–	.46‡	**.51**	.48	.44‡	.45‡	.35‡
DCU	**.65‡**	**.59‡**	**.58‡**	**.57‡**	**.56‡**	**.55‡**	**.54‡**	–	.50	.48	.48	.45‡	.36‡
DCU-FDA	**.63‡**	**.57‡**	**.64‡**	**.60‡**	**.59‡**	**.56‡**	.49	.50	–	**.53⋆**	.49	.42‡	.32‡
DCU-OKITA	**.66‡**	**.62‡**	**.62‡**	**.63‡**	**.60‡**	**.60‡**	**.52**	**.52**	.47⋆	–	.50	.47†	.36‡
CU-ZEMAN	**.67‡**	**.62‡**	**.62‡**	**.61‡**	**.57‡**	**.60‡**	**.56‡**	**.52**	**.51**	.50	–	.46‡	.40‡
JHU	**.68‡**	**.63‡**	**.62‡**	**.66‡**	**.64‡**	**.63‡**	**.55‡**	**.55‡**	**.58‡**	**.53†**	**.54‡**	–	.37‡
SHEF-WPROA	**.77‡**	**.69‡**	**.74‡**	**.71‡**	**.68‡**	**.68‡**	**.65‡**	**.64‡**	**.68‡**	**.64‡**	**.60‡**	**.63‡**	–
score	.63	.58	.57	.56	.53	.52	.49	.47	.47	.45	.44	.41	.32
rank	1	2-4	2-4	3-4	5-6	5-6	7-8	7-9	8-10	9-11	10-11	12	13

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

	ONLINE-B	CMU	ONLINE-A	ONLINE-G	PROMT	QCRI-MES	UCAM-MULTIFRONTEND	BALAGUR	MES-QCRI	UEDIN	OMNIFLUENT-UNCNSTR	LIA	OMNIFLUENT-CNSTR	UMD	CU-KAREL	COMMERCIAL-3	UEDIN-SYNTAX	JHU	CU-ZEMAN
ONLINE-B	–	.40‡	.42‡	.41‡	.37‡	.37‡	.41‡	.33‡	.33‡	.37‡	.33‡	.33‡	.35‡	.38‡	.34‡	.33‡	.29‡	.28‡	.14‡
CMU	**.60‡**	–	.50	.46†	.43‡	.47†	.42‡	.42‡	.39‡	.43‡	.41‡	.41‡	.40‡	.38‡	.36‡	.30‡	.30‡	.29‡	.17‡
ONLINE-A	**.58‡**	.50	–	.50	**.51**	.43‡	.47⋆	.44‡	.40‡	.41‡	.43‡	.38‡	.40‡	.38‡	.38‡	.39‡	.34‡	.30‡	.19‡
ONLINE-G	**.59‡**	**.54†**	.50	–	**.55†**	.50	**.51**	.48	.42‡	.41‡	.44‡	.43‡	.46†	.40‡	.44‡	.36‡	.34‡	.33‡	.19‡
PROMT	**.63‡**	**.57‡**	.49	.45†	–	.43‡	.47†	.43‡	.47†	.47†	.43‡	.39‡	.44‡	.43‡	.37‡	.41‡	.40‡	.38‡	.25‡
QCRI-MES	**.63‡**	**.53†**	**.57‡**	.50	**.57‡**	–	.48	.46†	.47⋆	.45‡	.43‡	.45‡	.45‡	.38‡	.42‡	.37‡	.33‡	.40‡	.19‡
UCAM-MULTIFRONTEND	**.59‡**	**.58‡**	**.53⋆**	.49	**.53†**	**.52**	–	.47†	.48	.46‡	.46‡	.42‡	.45‡	.46‡	.45‡	.40‡	.39‡	.33‡	.17‡
BALAGUR	**.67‡**	**.58‡**	**.56‡**	**.52**	**.57‡**	**.54†**	**.53†**	–	.47†	.49	.45‡	**.53⋆**	.40‡	.44‡	.44‡	.41‡	.36‡	.33‡	.23‡
MES-QCRI	**.67‡**	**.61‡**	**.60‡**	**.58‡**	**.53†**	**.53⋆**	**.52**	**.53†**	–	.49	.47†	.47⋆	.43‡	.43‡	.44‡	.38‡	.42‡	.39‡	.17‡
UEDIN	**.63‡**	**.57‡**	**.59‡**	**.59‡**	**.53†**	**.55‡**	**.54‡**	**.51**	**.51**	–	.48	**.52**	.44‡	**.52**	.49	.42‡	.43‡	.35‡	.21‡
OMNIFLUENT-UNCNSTR	**.67‡**	**.59‡**	**.57‡**	**.56‡**	**.57‡**	**.57‡**	**.54‡**	**.55‡**	**.53†**	**.52**	–	**.51**	.46†	.48	.48	.44‡	.40‡	.39‡	.25‡
LIA	**.67‡**	**.59‡**	**.62‡**	**.57‡**	**.61‡**	**.55‡**	**.58‡**	.47⋆	**.53⋆**	.48	.49	–	**.51**	.49	.48	.50	.41‡	.39‡	.20‡
OMNIFLUENT-CNSTR	**.65‡**	**.60‡**	**.60‡**	**.54†**	**.56‡**	**.55‡**	**.55‡**	**.60‡**	**.57‡**	**.56‡**	**.54†**	.49	–	**.51**	.48	.47⋆	.40‡	.40‡	.25‡
UMD	**.62‡**	**.62‡**	**.62‡**	**.60‡**	**.57‡**	**.62‡**	**.54‡**	**.56‡**	**.57‡**	.48	**.52**	**.51**	.49	–	**.53†**	.42‡	.46‡	.42‡	.19‡
CU-KAREL	**.66‡**	**.64‡**	**.62‡**	**.56‡**	**.63‡**	**.58‡**	**.55‡**	**.56‡**	**.56‡**	**.51**	**.52**	**.52**	**.52**	.47†	–	.44‡	.40‡	.47⋆	.24‡
COMMERCIAL-3	**.67‡**	**.70‡**	**.61‡**	**.64‡**	**.59‡**	**.63‡**	**.60‡**	**.59‡**	**.62‡**	**.58‡**	**.56‡**	.50	**.53⋆**	**.58‡**	**.56‡**	–	**.51**	.44‡	.32‡
UEDIN-SYNTAX	**.71‡**	**.70‡**	**.66‡**	**.66‡**	**.60‡**	**.67‡**	**.61‡**	**.64‡**	**.58‡**	**.57‡**	**.60‡**	**.59‡**	**.60‡**	**.54‡**	**.60‡**	.49	–	.45‡	.25‡
JHU	**.72‡**	**.71‡**	**.70‡**	**.67‡**	**.62‡**	**.60‡**	**.67‡**	**.67‡**	**.61‡**	**.65‡**	**.61‡**	**.61‡**	**.60‡**	**.58‡**	**.53⋆**	**.56‡**	**.55‡**	–	.24‡
CU-ZEMAN	**.86‡**	**.83‡**	**.81‡**	**.81‡**	**.75‡**	**.81‡**	**.83‡**	**.77‡**	**.83‡**	**.79‡**	**.75‡**	**.80‡**	**.75‡**	**.81‡**	**.76‡**	**.68‡**	**.75‡**	**.76‡**	–
score	.65	.60	.58	.56	.56	.55	.54	.52	.51	.50	.49	.49	.48	.48	.47	.43	.41	.39	.21
rank	1	2-3	2-3	4-6	4-6	5-7	5-7	8-9	8-10	9-11	10-12	11-14	12-15	12-15	13-15	16	17	18	19

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

	PROMT	ONLINE-B	CMU	ONLINE-G	ONLINE-A	UEDIN	QCRI-MES	CU-KAREL	MES-QCRI	JHU	COMMERCIAL-3	LIA	BALAGUR	CU-ZEMAN
PROMT	–	.44‡	.39‡	.47	.46⋆	.36‡	.37‡	.37‡	.32‡	.35‡	.28‡	.30‡	.32‡	.24‡
ONLINE-B	**.56‡**	–	.44‡	.41‡	.44†	.38‡	.37‡	.35‡	.33‡	.39‡	.33‡	.31‡	.35‡	.24‡
CMU	**.61‡**	**.56‡**	–	**.52**	.49	.47†	.43‡	.41‡	.39‡	.44‡	.44‡	.40‡	.35‡	.28‡
ONLINE-G	**.53**	**.59‡**	.48	–	.48	.50	.48	.46	.46⋆	.42‡	.38‡	.43‡	.38‡	.36‡
ONLINE-A	**.54⋆**	**.56†**	**.51**	**.52**	–	.47	.49	.49	.48	.44†	.38‡	.40‡	.40‡	.34‡
UEDIN	**.64‡**	**.62‡**	**.53†**	.50	**.53**	–	.49	.46†	.42‡	.39‡	.44‡	.41‡	.38‡	.29‡
QCRI-MES	**.63‡**	**.63‡**	**.57‡**	**.52**	**.51**	**.51**	–	.48	.45‡	.44‡	.42‡	.39‡	.40‡	.29‡
CU-KAREL	**.63‡**	**.65‡**	**.59‡**	**.54**	**.51**	**.54†**	**.52**	–	.50	.46†	.43‡	.40‡	.42‡	.34‡
MES-QCRI	**.68‡**	**.67‡**	**.61‡**	**.54⋆**	**.52**	**.58‡**	**.55‡**	.50	–	.48⋆	.47‡	.43‡	.45‡	.34‡
JHU	**.65‡**	**.61‡**	**.56‡**	**.58‡**	**.56†**	**.61‡**	**.56‡**	**.54†**	**.52⋆**	–	**.51**	.44‡	.44‡	.33‡
COMMERCIAL-3	**.72‡**	**.67‡**	**.56‡**	**.62‡**	**.62‡**	**.56‡**	**.58‡**	**.57‡**	**.53‡**	.49	–	**.52**	.48	.44‡
LIA	**.70‡**	**.69‡**	**.60‡**	**.57‡**	**.60‡**	**.59‡**	**.61‡**	**.60‡**	**.57‡**	**.56‡**	.48	–	.47†	.41‡
BALAGUR	**.68‡**	**.65‡**	**.65‡**	**.62‡**	**.60‡**	**.62‡**	**.60‡**	**.58‡**	**.55‡**	**.56‡**	**.52**	**.53†**	–	.41‡
CU-ZEMAN	**.76‡**	**.76‡**	**.72‡**	**.64‡**	**.66‡**	**.71‡**	**.71‡**	**.66‡**	**.66‡**	**.67‡**	**.56‡**	**.59‡**	**.59‡**	–
score	.64	.62	.55	.54	.53	.53	.52	.49	.47	.46	.43	.42	.41	.33
rank	1	2	3-4	3-6	3-7	4-7	5-7	8	9-10	9-10	11-12	11-13	12-13	14

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