Findings of the 2011 Workshop on Statistical Machine Translation

Chris Callison-Burch
Center for Language and Speech Processing
Johns Hopkins University

Philipp Koehn
School of Informatics
University of Edinburgh

Christof Monz
Informatics Institute
University of Amsterdam

Omar F. Zaidan
Center for Language and Speech Processing
Johns Hopkins University

Abstract

This paper presents the results of the WMT11 shared tasks, which included a translation task, a system combination task, and a task for machine translation evaluation metrics. We conducted a large-scale manual evaluation of 148 machine translation systems and 41 system combination entries. We used the ranking of these systems to measure how strongly automatic metrics correlate with human judgments of translation quality for 21 evaluation metrics. This year featured a Haitian Creole to English task translating SMS messages sent to an emergency response service in the aftermath of the Haitian earthquake. We also conducted a pilot 'tunable metrics' task to test whether optimizing a fixed system to different metrics would result in perceptibly different translation quality.

1 Introduction

This paper presents the results of the shared tasks of the Workshop on statistical Machine Translation (WMT), which was held at EMNLP 2011. This workshop builds on five previous WMT workshops (Koehn and Monz, 2006; Callison-Burch et al., 2007; Callison-Burch et al., 2008; Callison-Burch et al., 2009; Callison-Burch et al., 2010). The workshops feature three shared tasks: a translation task between English and other languages, a task to combine the output of multiple machine translation systems, and a task to predict human judgments of translation quality using automatic evaluation metrics. The performance for each of these shared tasks is determined through a comprehensive human evaluation. There were a two additions to this year's workshop that were not part of previous workshops:

- **Haitian Creole featured task** – In addition to translation between European language pairs, we featured a new translation task: translating Haitian Creole SMS messages that were sent to an emergency response hotline in the immediate aftermath of the 2010 Haitian earthquake. The goal of this task is to encourage researchers to focus on challenges that may arise in future humanitarian crises. We invited Will Lewis, Rob Munro and Stephan Vogel to publish a paper about their experience developing translation technology in response to the crisis (Lewis et al., 2011). They provided the data used in the Haitian Creole featured translation task. We hope that the introduction of this new dataset will provide a testbed for dealing with low resource languages and the informal language usage found in SMS messages.

- **Tunable metric shared task** – We conducted a pilot of a new shared task to use evaluation metrics to tune the parameters of a machine translation system. Although previous workshops have shown evaluation metrics other than BLEU are more strongly correlated with human judgments when ranking outputs from multiple systems, BLEU remains widely used by system developers to optimize their system parameters. We challenged metric developers to tune the parameters of a fixed system, to see if their metrics would lead to perceptibly better translation quality for the system's resulting output.

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 with previous workshops, all of the data, translations, and collected human judgments are publicly available.[1] We hope these datasets form a valuable resource for research into statistical machine translation, system combination, and automatic evaluation of translation quality.

2 Overview of the Shared Translation and System Combination Tasks

The recurring task of the workshop examines translation between English and four other languages: German, Spanish, French, and Czech. We created a test set for each language pair by translating newspaper articles. We additionally provided training data and two baseline systems.

2.1 Test data

The test data for this year's task was created by hiring people to translate news articles that were drawn from a variety of sources from early December 2010. A total of 110 articles were selected, in roughly equal amounts from a variety of Czech, English, French, German, and Spanish news sites:[2]

Czech: aktualne.cz (4), Novinky.cz (7), iHNed.cz (4), iDNES.cz (4)

French: Canoe (5), Le Devoir (5), Le Monde (5), Les Echos (5), Liberation (5)

Spanish: ABC.es (6), Cinco Dias (6), El Periodico (6), Milenio (6), Noroeste (7)

English: Economist (4), Los Angeles Times (6), New York Times (4), Washington Post (4)

German: FAZ (3), Frankfurter Rundschau (2), Financial Times Deutschland (3), Der Spiegel (5), Süddeutsche Zeitung (3)

The translations were created by the professional translation agency CEET.[3] All of the translations were done directly, and not via an intermediate language.

Although the translations were done professionally, in some cases errors still cropped up. For instance, in parts of the English-French translations, some of the English source remains in the French reference as if the translator forgot to delete it.

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 statistics about the training materials are given in Figure 1.

2.3 Baseline systems

To lower the barrier of entry for newcomers to the field, we provided two open source toolkits for phrase-based and parsing-based statistical machine translation (Koehn et al., 2007; Li et al., 2010).

2.4 Submitted systems

We received submissions from 56 groups across 37 institutions, as listed in Tables 1, 2 and 3. We also included two commercial off-the-shelf MT systems, two online statistical MT systems, and five online rule-based MT systems. (Not all systems supported all language pairs.) We note that these nine companies did not submit entries themselves, and are therefore anonymized in this paper. Rather, their entries were created by translating the test data via their web interfaces.[4] The data used to construct these systems is not subject to the same constraints as the shared task participants. It is possible that part of the reference translations that were taken from online news sites could have been included in the online systems' models, for instance. We therefore categorize all commercial systems as unconstrained when evaluating the results.

2.5 System combination

In total, we had 148 primary system entries (including the 46 entries crawled from online sources), and 60 contrastive entries. These were made available to

[1] `http://statmt.org/wmt11/results.html`

[2] 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.

[3] `http://www.ceet.eu/`

[4] We would like to thank Ondřej Bojar for harvesting the commercial entries (2), Christian Federmann for the statistical MT entries (14), and Hervé Saint-Amand for the rule-based MT entries (30)!

Europarl Training Corpus

	Spanish ↔ English		French ↔ English		German ↔ English		Czech ↔ English	
Sentences	1,786,594		1,825,077		1,739,154		462,351	
Words	51,551,370	49,411,045	54,568,499	50,551,047	45,607,269	47,978,832	10,573,983	12,296,772
Distinct words	171,174	113,655	137,034	114,487	362,563	111,934	152,788	56,095

News Commentary Training Corpus

	Spanish ↔ English		French ↔ English		German ↔ English		Czech ↔ English	
Sentences	132,571		115,562		136,227		122,754	
Words	3,739,293	3,285,305	3,290,280	2,866,929	3,401,766	3,309,619	2,658,688	2,951,357
Distinct words	73,906	53,699	59,911	50,323	120,397	53,921	130,685	50,457

United Nations Training Corpus

	Spanish ↔ English		French ↔ English	
Sentences	10,662,993		12,317,600	
Words	348,587,865	304,724,768	393,499,429	344,026,111
Distinct words	578,599	564,489	621,721	729,233

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 Training Corpus

	Czech ↔ English	
Sentences	7,227,409	
Words	72,993,427	84,856,749
Distinct words	1,088,642	522,770

Europarl Language Model Data

	English	Spanish	French	German	Czech
Sentence	2,032,006	1,942,761	2,002,266	1,985,560	479,636
Words	54,720,731	55,105,358	57,860,307	48,648,697	10,770,230
Distinct words	119,315	176,896	141,742	376,128	154,129

News Language Model Data

	English	Spanish	French	German	Czech
Sentence	30,888,595	3,416,184	11,767,048	17,474,133	12,333,268
Words	777,425,517	107,088,554	302,161,808	289,171,939	216,692,489
Distinct words	2,020,549	595,681	1,250,259	3,091,700	2,068,056

News Test Set

	English	Spanish	French	German	Czech
Sentences	3003				
Words	75,762	79,710	85,999	73,729	65,427
Distinct words	10,088	11,989	11,584	14,345	16,922

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	Participant
ALACANT	University of Alicante (Sánchez-Cartagena et al., 2011)
CEU-UPV	CEU University Cardenal Herrera & Polytechnic University of Valencia (Zamora-Martinez and Castro-Bleda, 2011)
CMU-DENKOWSKI	Carnegie Mellon University - Denkowski (Denkowski and Lavie, 2011b)
CMU-DYER	Carnegie Mellon University - Dyer (Dyer et al., 2011)
CMU-HANNEMAN	Carnegie Mellon University - Hanneman (Hanneman and Lavie, 2011)
COPENHAGEN	Copenhagen Business School
CST	Centre for Language Technology @ Copenhagen University (Rishøj and Søgaard, 2011)
CU-BOJAR	Charles University - Bojar (Mareček et al., 2011)
CU-MARECEK	Charles University - Mareček (Mareček et al., 2011)
CU-POPEL	Charles University - Popel (Popel et al., 2011)
CU-TAMCHYNA	Charles University - Tamchyna (Bojar and Tamchyna, 2011)
CU-ZEMAN	Charles University - Zeman (Zeman, 2011)
DFKI-FEDERMANN	Deutsches Forschungszentrum für Künstliche Intelligenz - Federmann (Federmann and Hunsicker, 2011)
DFKI-XU	Deutsches Forschungszentrum für Künstliche Intelligenz - Xu (Xu et al., 2011b)
HYDERABAD	IIIT-Hyderabad
ILLC-UVA	Institute for Logic, Language and Computation @ University of Amsterdam (Khalilov and Sima'an, 2011)
JHU	Johns Hopkins University (Weese et al., 2011)
KIT	Karlsruhe Institute of Technology (Herrmann et al., 2011)
KOC	Koc University (Bicici and Yuret, 2011)
LATL-GENEVA	Language Technology Laboratory @ University of Geneva (Wehrli et al., 2009)
LIA-LIG	Laboratoire Informatique d'Avignon @ The University of Avignon & Laboratoire d'Informatique de Grenoble @ University of Grenoble (Potet et al., 2011)
LIMSI	LIMSI (Allauzen et al., 2011)
LINGUATEC	Linguatec Language Technologies (Aleksic and Thurmair, 2011)
LIU	Linköping University (Holmqvist et al., 2011)
LIUM	University of Le Mans (Schwenk et al., 2011)
PROMT	ProMT
RWTH-FREITAG	RWTH Aachen - Freitag (Huck et al., 2011)
RWTH-HUCK	RWTH Aachen - Huck (Huck et al., 2011)
RWTH-WUEBKER	RWTH Aachen - Wübker (Huck et al., 2011)
SYSTRAN	SYSTRAN
UEDIN	University of Edinburgh (Koehn et al., 2007)
UFAL-UM	Charles University and University of Malta (Corbí-Bellot et al., 2005)
UOW	University of Wolverhampton (Aziz et al., 2011)
UPM	Technical University of Madrid (López-Ludeña and San-Segundo, 2011)
UPPSALA	Uppsala University (Koehn et al., 2007)
UPPSALA-FBK	Uppsala University & Fondazione Bruno Kessler (Hardmeier et al., 2011)
ONLINE-[A,B]	two online statistical machine translation systems
RBMT-[1–5]	five online rule-based machine translation systems
COMMERCIAL-[1,2]	two commercial machine translation systems

Table 1: Participants in the shared translation task (European language pairs; individual system track). Not all teams participated in all language pairs. The translations from commercial and online systems were crawled by us, not submitted by the respective companies, and are therefore anonymized.

ID	Participant
BBN-COMBO	Raytheon BBN Technologies (Rosti et al., 2011)
CMU-HEAFIELD-COMBO	Carnegie Mellon University (Heafield and Lavie, 2011)
JHU-COMBO	Johns Hopkins University (Xu et al., 2011a)
KOC-COMBO	Koc University (Bicici and Yuret, 2011)
LIUM-COMBO	University of Le Mans (Barrault, 2011)
QUAERO-COMBO	Quaero Project* (Freitag et al., 2011)
RWTH-LEUSCH-COMBO	RWTH Aachen (Leusch et al., 2011)
UOW-COMBO	University of Wolverhampton (Specia et al., 2010)
UPV-PRHLT-COMBO	Polytechnic University of Valencia (González-Rubio and Casacuberta, 2011)
UZH-COMBO	University of Zurich (Sennrich, 2011)

Table 2: Participants in the shared system combination task. Not all teams participated in all language pairs.
* The Quaero Project entry combined outputs they received directly from LIMSI, KIT, SYSTRAN, and RWTH.

participants in the system combination shared task. Continuing our practice from last year's workshop, we separated the test set into a tuning set and a final held-out test set for system combinations. The tuning portion was distributed to system combination participants along with reference translations, to aid them set any system parameters.

In the European language pairs, the tuning set consisted of 1,003 segments taken from 37 documents, whereas the test set consisted of 2,000 segments taken from 73 documents. In the Haitian Creole task, the split was 674 segments for tuning and 600 for testing.

Table 2 lists the 10 participants in the system combination task.

3 Featured Translation Task

The featured translation task of WMT11 was to translate Haitian Creole SMS messages into English. These text messages were sent by people in Haiti in the aftermath of the January 2010 earthquake. In the wake of the earthquake, much of the country's conventional emergency response services failed. Since cell phone towers remained standing after the earthquake, text messages were a viable mode of communication. Munro (2010) describes how a text-message-based emergency reporting system was set up by a consortium of volunteer organizations named "Mission 4636" after a free SMS short code telephone number that they established. The SMS messages were routed to a system for reporting trapped people and other emergencies. Search and rescue teams within Haiti, including the US Military, recognized the quantity and reliability of actionable information in these messages and used them to provide aid.

The majority of the SMS messages were written in Haitian Creole, which was not spoken by most of first responders deployed from overseas. A distributed, online translation effort was established, drawing volunteers from Haitian Creole- and French-speaking communities around the world. The volunteers not only translated messages, but also categorized them and pinpointed them on a map.[5] Collaborating online, they employed their local knowledge of locations, regional slang, abbreviations and spelling variants to process more than 40,000 messages in the first six weeks alone. First responders indicated that this volunteer effort helped to save hundreds of lives and helped direct the first food and aid to tens of thousands. Secretary of State Clinton described one success of the Mission 4636 program:"The technology community has set up interactive maps to help us identify needs and target resources. And on Monday, a seven-year-old girl and two women were pulled from the rubble of a collapsed supermarket by an American search-and-rescue team after they sent a text message calling for help." Ushahidi@Tufts described another:"The World Food Program delivered food to an informal camp of 2500 people, having yet to receive food or water, in Diquini to a location that 4636 had identi-

[5] A detailed map of Haiti was created by a crowdsourcing effort in the aftermath of the earthquake (Lacey-Hall, 2011).

ID	Participant
BM-I2R	Barcelona Media & Institute for Infocomm Research (Costa-jussà and Banchs, 2011)
CMU-DENKOWSKI	Carnegie Mellon University - Denkowski (Denkowski and Lavie, 2011b)
CMU-HEWAVITHARANA	Carnegie Mellon University - Hewavitharana (Hewavitharana et al., 2011)
HYDERABAD	IIIT-Hyderabad
JHU	Johns Hopkins University (Weese et al., 2011)
KOC	Koc University (Bicici and Yuret, 2011)
LIU	Linköping University (Stymne, 2011)
UMD-EIDELMAN	University of Maryland - Eidelman (Eidelman et al., 2011)
UMD-HU	University of Maryland - Hu (Hu et al., 2011)
UPPSALA	Uppsala University (Hardmeier et al., 2011)

Table 3: Participants in the featured translation task (Haitian Creole SMS into English; individual system track). Not all teams participated in both the 'Clean' and 'Raw' tracks.

fied for them."

In parallel with Rob Munro's crowdsourcing translation efforts, the Microsoft Translator team developed a Haitian Creole statistical machine translation engine from scratch in a compressed timeframe (Lewis, 2010). Despite the impressive number of translations completed by volunteers, machine translation was viewed as a potentially useful tool for higher volume applications or to provide translations of English medical documents into Haitian Creole. The Microsoft Translator team quickly assembled parallel data from a number of sources, including Mission 4636 and from the archives of Carnegie Mellon's DIPLOMAT project (Frederking et al., 1997). Through a series of rapid prototyping efforts, the team improved their system to deal with non-standard orthography, reduced pronouns, and SMS shorthand. They deployed a functional translation system to relief workers in the field in less than 5 days – impressive even when measured against previous rapid MT development efforts like DARPA's surprise language exercise (Oard, 2003; Oard and Och, 2003).

We were inspired by the efforts of Rob Munro and Will Lewis on translating Haitian Creole in the aftermath of the disaster, so we worked with them to create a featured task at WMT11. We thank them for generously sharing the data they assembled in their own efforts. We invited Rob Munro, Will Lewis, and Stephan Vogel to speak at the workshop on the topic of developing translation technology for future crises, and they recorded their thoughts in an invited publication (Lewis et al., 2011).

3.1 Haitian Creole Data

For the WMT11 featured translation task, we anonymized the SMS Haitian Creole messages along with the translations that the Mission 4636 volunteers created. Examples of these messages are given in Table 4. The goal of anonymizing the SMS data was so that it may be shared with researchers who are developing translation and mapping technologies to support future emergency relief efforts and social development. We ask that any researcher working with these messages to be aware that they are actual communications sent by people in need in a time of crisis. Researchers who use this data are asked to be cognizant of the following:

- Some messages may be distressing in content.
- The people who sent the messages (and who are discussed in them) were victims of a natural disaster and a humanitarian crisis. Please treat the messages with the appropriate respect for these individuals.
- The primary motivation for using this data should be to understand how we can better respond to future crises.

Participants who received the Haitian Creole data for WMT11 were given anonymization guidelines

mwen se [FIRSTNAME] mwen gen twaset ki mouri mwen mande nou ed pou nou edem map tan repons	I am [FIRSTNAME], I have three sisters who have died. I ask help for us, I await your response.
Ki kote yap bay manje	Where are they giving out food?
Eske lekol kolej marie anne kraze?mesi	Was the College Marie Anne school destroyed? Thank you.
Nou pa ka anpeche moustik yo mòde nou paske yo anpil.	We can't prevent the mosquitoes from biting because there are so many.
tanpri kèm ap kase mwen pa ka pran nouvel manmanm.	Please heart is breaking because I have no news of my mother.
4636:Opital Medesen san Fwontiè delmas 19 la fèmen. Opital sen lwi gonzag nan delma 33 pran an chaj gratwitman tout moun ki malad ou blese	4636: The Doctors without Borders Hospital in Delmas 19 is closed. The Saint Louis Gonzaga hospital in Delmas 33 is taking in sick and wounded people for free
Mwen résévoua mesaj nou yo 5 sou 5 men mwen ta vle di yon bagay kilè e koman nap kapab fèm jwin èd sa yo pou moune b la kay mwen ki sinistwé adrès la sé	I received your message 5/5 but I would like to ask one thing when and how will you be able to get the aid to me for the people around my house who are victims of the earthquake? The address is
Sil vous plait map chehe [LASTNAME][FIRSTNAME].di yo relem nan [PHONENUMBER].mwen se [LASTNAME] [FIRSTNAME]	I'm looking for [LASTNAME][FIRSTNAME]. Tell him to call me at [PHONENUMBER] I am [LASTNAME] [FIRSTNAME]
Bonswa mwen rele [FIRSTNAME] [LASTNAME] kay mwen krase mwen pagin anyin poum mange ak fanmi-m tampri di yon mo pou mwen fem jwen yon tante tou ak mange. .mrete n	Hello my name is [FIRSTNAME] [LASTNAME]my house fell down, I've had nothing to eat and I'm hungry. Please help me find food. I live
Mwen viktim kay mwen kraze èskem ka ale sendomeng mwen gen paspò	I'm a victim. My home has been destroyed. Am I allowed to go to the Dominican Republic? I have a Passport.
KISAM DWE FE LEGEN REPLIK,ESKE MOUN SAINT MARC AP JWENN REPLIK.	What should I do when there is an aftershock? Will the people of Saint Marc have aftershocks?
MWEN SE YON JEN ETIDYAN AN ASYANS ENFOMATIK KI PASE ANPIL MIZE NAN TRANBLEMAN DE TE 12 JANVYE A TOUT FANMIM FIN MOURI MWEN SANTIM SEL MWEN TE VLE ALE VIV	I'm a young student in computer science, who has suffered a lot during and after the earthquake of January 12th. All my family has died and I feel alone. I wanted to go live.
Mw rele [FIRSTNAME], mw fè mason epi mw abite laplèn. Yo dim minustah ap bay djob mason ki kote pou mw ta pase si mw ta vle jwenn nan djob sa yo.	My name is [FIRSTNAME], I'm a construction worker and I live in La Plaine. I heard that the MINUSTAH was giving jobs to construction workers. What do I have to go to find one of these jobs?
Souple mande lapolis pou fe on ti pase nan magloire ambroise prolonge zone muler ak cadet jeremie ginyin jen gason ki ap pase nan zone sa yo e ki agresi	please ask the police to go to magloire ambroise going towards the "muler" area and cadet jeremie because there are very aggressive young men in these areas
KIBO MOUN KA JWENN MANJE POU YO MANJE ANDEYO KAPITAL PASKE DEPI 12 JANVYE YO VOYE MANJE POU PEP LA MEN NOU PA JANM JWENN ANYEN. NAP MOURI AK GRANGOU	Where can people get food to eat outside of the capital because since January 12th, they've sent food for the people but we never received anything. We are dying of hunger
Mwen se [FIRSTNAME][LASTNAME] mwen nan aken mwen se yon jèn ki ansent mwen te genyen yon paran ki tap ede li mouri pòtoprens, mwen pral akouye nan kòmansman feviye	I am [FIRSTNAME][LASTNAME] I am in Aquin I am a pregnant young person I had a parent who was helping me, she died in Port-au-Prince, I'm going to give birth at the start of February

Table 4: Examples of some of the Haitian Creole SMS messages that were sent to the 4636 short code along with their translations into English. Translations were done by volunteers who wanted to help with the relief effort. Prior to being distributed, the messages were anonymized to remove names, phone numbers, email addresses, etc. The anonymization guidelines specified that addresses be retained to facilitate work on mapping technologies.

Training set	Parallel sentences	Words per lang
In-domain SMS data	17,192	35k
Medical domain	1,619	10k
Newswire domain	13,517	30k
Glossary	35,728	85k
Wikipedia parallel sentence	8,476	90k
Wikipedia named entities	10,499	25k
The bible	30,715	850k
Haitisurf dictionary	3,763	4k
Krengle dictionary	1,687	3k
Krengle sentences	658	3k

Table 5: Training data for the Haitian Creole-English featured translation task. The in-domain SMS data consists primarily of raw (noisy) SMS data. The in-domain data was provided by Mission 4636. The other data is out-of-domain. It comes courtesy of Carnegie Mellon University, Microsoft Research, Haitisurf.com, and Krengle.net.

alongside the SMS data. The WMT organizers requested that if they discovered messages with incorrect or incomplete anonymization, that they notify us and correct the anonymization using the version control repository.

To define the shared translation task, we divided the SMS messages into an in-domain training set, along with designated dev, devtest, and test sets. We coordinated with Microsoft and CMU to make available additional out-of-domain parallel corpora. Details of the data are given in Table 5. In addition to this data, participants in the featured task were allowed to use any of the data provided in the standard translation task, as well as linguistic tools such as taggers, parsers, or morphological analyzers.

3.2 Clean and Raw Test Data

We provided two sets of testing and development data. Participants used their systems to translate two test sets consisting of 1,274 unseen Haitian Creole SMS messages. One of the test sets contains the "raw" SMS messages as they were sent, and the other contains messages that were cleaned up by human post-editors. The English side is the same in both cases, and the only difference is the Haitian Creole input sentences.

The post-editors were Haitian Creole language informants hired by Microsoft Research. They provided a number of corrections to the SMS messages, including expanding SMS shorthands, correcting spelling/grammar/capitalization, restoring diacritics that were left out of the original message, and cleaning up accented characters that were lost when the message was transmitted in the wrong encoding.

Original Haitian Creole messages:

> Sil vou plé éde mwen avek moun ki viktim yo nan tranbleman de té a,ki kité potoprins ki vini nan provins- mwen ede ak ti kob mwen te ginyin kounié a
> 4636: Manje vin pi che nan PaP apre tranbleman te-a. mamit diri ap van'n 250gd kounye, sete 200gd avan. Mayi-a 125gd, avan sete 100gd

Edited Haitian Creole messages:

> Silvouple ede mwen avèk moun ki viktim yo nan tranblemanntè a, ki kite Pòtoprens ki vini nan pwovens, mwen ede ak ti kòb mwen te genyen kounye a
> 4636: Manje vin pi chè nan PaP apre tranblemanntè a. Mamit diri ap vann 250gd kounye a, sete 200gd avan. Mayi-a 125gd, avan sete 100gd.

For the test and development sets the informants also edited the English translations. For instance, there were cases where the original crowdsourced translation summarized the content of the message instead of translating it, instances where parts of the source were omitted, and where explanatory notes were added. The editors improved the translations so that they were more suitable for machine translation, making them more literal, correcting disfluencies on the English side, and retranslating them when they were summaries.

Crowdsourced English translation:

> We are in the area of Petit Goave, we would like .... we need tents and medication for flu/colds...

Post-edited translation:

> We are in the area of Petit Goave, we would like to receive assistance, however,

it should not be the way I see the Minustah guys are handling the people. We need lots of tents and medication for flu/colds, and fever

The edited English is provided as the reference for both the "clean" and the "raw" sets, since we intend that distinction to refer to the form that the source language comes in, rather than the target language.

Tables 47 and 48 in the Appendix show a significant difference in the translation quality between the clean and the raw test sets. In most cases, systems' output for the raw condition was 4 BLEU points lower than for the clean condition. We believe that the difference in performance on the raw vs. cleaned test sets highlight the importance of handling noisy input data.

All of the in-domain training data is in the raw format. The original SMS messages are unaltered, and the translations are just as the volunteered provided them. In some cases, the original SMS messages are written in French or English instead of Haitian Creole, or contain a mixture of languages. It may be possible to further improve the quality of machine translation systems trained from this data by improving the quality of the data itself.

3.3 Goals and Challenges

The goals of the Haitian Creole to English translation task were:

- To focus researchers on the problems presented by low resource languages
- To provide a real-world data set consisting of SMS messages, which contain abbreviations, non-standard spelling, omitted diacritics, and other noisy character encodings
- To develop techniques for building translation systems that will be useful in future crises

There are many challenges in translating noisy data in a low resource language, and there are a variety of strategies that might be considered to attempt to tackle them. For instance:

- Automated cleaning of the raw (noisy) SMS data in the training set.
- Leveraging a larger French-English model to translate out of vocabulary Haitian words, by creating a mapping from Haitian words onto French.
- Incorporation of morphological and/or syntactic models to better cope with the low resource language pair.

It is our hope that by introducing this data as a shared challenge at WMT11 that we will establish a useful community resource so that researchers may explore these challenges and publish about them in the future.

4 Human Evaluation

As with past workshops, we placed greater emphasis on the human evaluation than on the automatic evaluation metric scores. It is our contention that automatic measures are an imperfect substitute for human assessment of translation quality. Therefore, we define the manual evaluation to be primary, and use the human judgments to validate automatic metrics.

Manual evaluation is time consuming, and it requires a large effort to conduct on the scale of our workshop. We distributed the workload across a number of people, including shared-task participants, interested volunteers, and a small number of paid annotators (recruited by the participating sites). More than 130 people participated in the manual evaluation, with 91 people putting in more than an hour's worth of effort, and 29 putting in more than four hours. There was a collective total of 361 hours of labor.

We asked annotators to evaluate system outputs by ranking translated sentences relative to each other. This was our official determinant of translation quality. The total number of judgments collected for the different ranking tasks is given in Table 6.

We performed the manual evaluation of the individual systems separately from the manual evaluation of the system combination entries, rather than comparing them directly against each other. Last year's results made it clear that there is a large (expected) gap in performance between the two groups. This year, we opted to reduce the number of pairwise

comparisons with the hope that we would be more likely to find statistically significant differences between the systems in the same groups. To that same end, we also eliminated the editing/acceptability task that was featured in last year's evaluation, instead we had annotators focus solely on the system ranking task.

4.1 Ranking translations of sentences

Ranking translations relative to each other is a reasonably intuitive task. We therefore kept the instructions simple:

> *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).*

Each screen for this task involved judging translations of three consecutive source segments. For each source segment, the annotator was shown the outputs of five submissions, and asked to rank them.

With the exception of a few tasks in the system combination track, there were many more than 5 systems participating in any given task—up to 23 for the English-German individual systems track. Rather than attempting to get a complete ordering over the systems, we instead relied on random selection and a reasonably large sample size to make the comparisons fair.

We use the collected rank labels 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 or equal to other systems. Specifically, each block in which A appears includes four implicit pairwise comparisons (against the other presented systems). A is rewarded once for each of the four comparisons in which A wins or ties. A's score is the number of such winning (or tying) pairwise comparisons, divided by the total number of pairwise comparisons involving A.

The system scores are reported in Section 5. Appendix A provides detailed tables that contain pairwise **head-to-head** comparisons between pairs of systems.

4.2 Inter- and Intra-annotator agreement in the ranking task

We were interested in determining the inter- and intra-annotator agreement for the ranking task, since a reasonable degree of agreement must exist to support our process as a valid evaluation setup. To ensure we had enough data to measure agreement, we purposely designed the sampling of source segments and translations shown to annotators in a way that ensured some items would be repeated, both within the screens completed by an individual annotator, and across screens completed by different annotators.

We did so by ensuring that 10% of the generated screens are exact repetitions of previously generated screen within the same batch of screens. Furthermore, even within the other 90%, we ensured that a source segment appearing in one screen appears again in two more screens (though with different system outputs). Those two details, intentional repetition of source sentences and intentional repetition of system outputs, ensured we had enough data to compute meaningful inter- and intra-annotator agreement rates.

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 $P(E)$. Note also that κ has a value of at most 1 (and could possibly be negative), with higher rates of agreement resulting in higher κ.

The above definition of κ is actually used by several definitions of agreement measures, which differ in how $P(A)$ and $P(E)$ are computed.

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

	Inividual System Track			System Combination Track		
Language Pair	# Systems	Label Count	Labels per System	# Systems	Label Count	Labels per System
Czech-English	8	2,490	276.7	4	1,305	261.0
English-Czech	10	8,985	816.8	2	2,700	900.0
German-English	20	4,620	220.0	8	1,950	216.7
English-German	22	6,540	284.4	4	2,205	441.0
Spanish-English	15	2,850	178.1	6	2,115	302.1
English-Spanish	15	5,595	349.7	4	3,000	600.0
French-English	18	3,540	186.3	6	1,500	214.3
English-French	17	4,590	255.0	2	900	300.0
Haitian (Clean)-English	9	3,360	336.0	3	1,200	300.0
Haitian (Raw)-English	6	1,875	267.9	2	900	300.0
Urdu-English (tunable metrics task)	8	3,165	351.7	N/A	N/A	N/A
Overall	**148**	**47,610**	**299.4**	**41**	**17,775**	**348.5**

Table 6: A summary of the WMT11 ranking task, showing the number of systems and number of labels collected in each of the individual and system combination tracks. The system count does not include the reference translation, which was included in the evaluation, and so a value under "Labels per System" can be obtained only after adding 1 to the system count, before dividing the label count (e.g. in German-English, $4,620/21 = 220.0$).

which annotators agree, in the context of pairwise comparisons. $P(A)$ is computed similarly for *intra*-annotator agreement (i.e. self-consistency), but over pairwise comparisons that were annotated more than once by a *single* annotator.

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

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

Note that each of the three probabilities in $P(E)$'s definition are squared to reflect the fact that we are considering the chance that *two* annotators would agree by chance. Each of these probabilities is computed empirically, by observing how often annotators actually rank two systems as being tied. We note here that this empirical computation is a departure from previous years' analyses, where we had assumed that the three categories are equally likely (yielding $P(E) = \frac{1}{9} + \frac{1}{9} + \frac{1}{9} = \frac{1}{3}$). We believe that this is a more principled approach, which faithfully reflects the motivation of accounting for $P(E)$ in the first place.[6]

Table 7 gives κ values for inter-annotator and intra-annotator agreement across the various evaluation tasks. These give an indication of how often different judges agree, and how often single judges are consistent for repeated judgments, respectively.

There are some general and expected trends that can be seen in this table. First of all, intra-annotator agreement is higher than inter-annotator agreement. Second, reference translations are noticeably better than other system outputs, which means that annotators have an artificially high level of agreement on pairwise comparisons that include a reference translation. For this reason, we also report the agreement levels when such comparisons are excluded.

The exact interpretation of the kappa coefficient is difficult, but according to Landis and Koch (1977), $0-0.2$ is slight, $0.2-0.4$ is fair, $0.4-0.6$ is moderate, $0.6-0.8$ is substantial, and $0.8-1.0$ is almost perfect. Based on these interpretations, the agreement for sentence-level ranking is moderate to substantial for most tasks.

[6]Even if we wanted to assume a "random clicker" model, setting $P(E) = \frac{1}{3}$ is still not entirely correct. Given that annotators rank five outputs at once, $P(A = B) = \frac{1}{5}$, not $\frac{1}{3}$, since there are only five (out of 25) label pairs that satisfy $A = B$. Working this back into $P(E)$'s definition, we have $P(A > B) = P(A < B) = \frac{2}{5}$, and therefore $P(E) = 0.36$ rather than 0.333.

INTER-ANNOTATOR AGREEMENT (I.E. ACROSS ANNOTATORS)

	ALL COMPARISONS			NO REF COMPARISONS		
	$P(A)$	$P(E)$	κ	$P(A)$	$P(E)$	κ
European languages, individual systems	0.601	0.362	0.375	0.561	0.355	0.320
European languages, system combinations	0.671	0.335	0.505	0.598	0.342	0.389
Haitian-English, individual systems	0.691	0.362	0.516	0.639	0.350	0.446
Haitian-English, system combinations	0.761	0.358	0.628	0.674	0.335	0.509
Tunable metrics task (Urdu-English)	0.692	0.337	0.535	0.641	0.363	0.437
WMT10 (European languages, all systems)	0.658	0.374	0.454	0.626	0.367	0.409

INTRA-ANNOTATOR AGREEMENT (I.E. SELF-CONSISTENCY)

	ALL COMPARISONS			NO REF COMPARISONS		
	$P(A)$	$P(E)$	κ	$P(A)$	$P(E)$	κ
European languages, individual systems	0.722	0.362	0.564	0.685	0.355	0.512
European languages, system combinations	0.787	0.335	0.680	0.717	0.342	0.571
Haitian-English, individual systems	0.763	0.362	0.628	0.700	0.350	0.539
Haitian-English, system combinations	0.882	0.358	0.816	0.784	0.335	0.675
Tunable metrics task (Urdu-English)	0.857	0.337	0.784	0.856	0.363	0.774
WMT10 (European languages, all systems)	0.755	0.374	0.609	0.734	0.367	0.580

Table 7: Inter- and intra-annotator agreement rates, for the various manual evaluation tracks of WMT11. See Tables 49 and 50 below for a detailed breakdown by language pair.

However, one result that is of concern is that agreement rates are noticeably lower for European language pairs, in particular for the individual systems track. When excluding reference comparisons, the inter- and intra-annotator agreement levels are 0.320 and 0.512, respectively. Not only are those numbers lower than for the other tasks, but they are also lower than last year's numbers, which were 0.409 and 0.580.

We investigated this result a bit deeper. Tables 49 and 50 in the Appendix break down the results further, by reporting agreement levels for each language pair. One observation is that the agreement level for some language pairs deviates in a non-trivial amount from the overall agreement rate.

Let us focus on inter-annotator agreement rates in the individual track (excluding reference comparisons), in the top right portion of Table 49. The overall κ is 0.320, but it ranges from 0.264 for German-English, to 0.477 for Spanish-English.

What distinguishes those two language pairs from each other? If we examine the results in Table 8, we see that Spanish-English had two very weak systems, which were likely easy for annotators to agree on comparisons involving them. (This is the converse of annotators agreeing more often on comparisons involving the reference.) English-French is similar in that regard, and it too has a relatively high agreement rate.

On the other hand, the participants in German-English formed a large pool of more closely-matched systems, where the gap separating the bottom system is not as pronounced. So it seems that the low agreement rates are indicative of a more competitive evaluation and more closely-matched systems.

5 Results of the Translation Tasks

We used the results of the manual evaluation to analyze the translation quality of the different systems that were submitted to the workshop. In our analysis, we aimed to address the following questions:

- Which systems produced the best translation quality for each language pair?
- Which of the systems that used only the provided training materials produced the best translation quality?

Czech-English
1023–1166 comparisons/system

System	C?	≥others
UEDIN ●⋆	Y	0.69
ONLINE-B ●	N	0.68
CU-BOJAR	N	0.60
JHU	N	0.57
UPPSALA	Y	0.57
SYSTRAN	N	0.51
CST	Y	0.47
CU-ZEMAN	Y	0.44

Spanish-English
583–833 comparisons/system

System	C?	≥others
ONLINE-B ●	N	0.72
ONLINE-A ●	N	0.72
KOC ⋆	Y	0.67
SYSTRAN ●	N	0.66
ALACANT ●	N	0.66
RBMT-1	N	0.63
RBMT-3	N	0.61
RBMT-2	N	0.60
RBMT-4	N	0.60
RBMT-5	N	0.51
UEDIN	Y	0.51
UPM	Y	0.50
UFAL-UM	Y	0.47
HYDERABAD	Y	0.17
CU-ZEMAN	Y	0.16

French-English
608–883 comparisons/system

System	C?	≥others
ONLINE-A ●	N	0.66
LIMSI ●⋆	Y+G	0.66
ONLINE-B ●	N	0.66
LIA-LIG	Y	0.64
KIT ●⋆	Y+G	0.64
LIUM	Y+G	0.63
CMU-DENKOWSKI ⋆	Y	0.62
JHU	Y+G	0.61
RWTH-HUCK	Y+G	0.58
RBMT-1 ●	N	0.58
CMU-HANNEMAN	Y+G	0.58
RBMT-3	N	0.55
SYSTRAN	N	0.54
RBMT-4	N	0.53
RBMT-2	N	0.52
UEDIN	Y	0.50
RBMT-5	N	0.45
CU-ZEMAN	Y	0.37

English-Czech
3126–3397 comparisons/system

System	C?	≥others
ONLINE-B ●	N	0.65
CU-BOJAR	N	0.64
CU-MARECEK ●	N	0.63
CU-TAMCHYNA	N	0.62
UEDIN ⋆	Y	0.59
CU-POPEL ⋆	Y	0.58
COMMERCIAL2	N	0.51
COMMERCIAL1	N	0.51
JHU	N	0.49
CU-ZEMAN	Y	0.43

English-Spanish
1300–1480 comparisons/system

System	C?	≥others
ONLINE-B ●	N	0.74
ONLINE-A ●	N	0.72
RBMT-3 ●	N	0.71
PROMT ●	N	0.70
CEU-UPV ⋆	Y	0.65
UEDIN ⋆	Y	0.64
UPPSALA ⋆	Y	0.61
RBMT-4	N	0.61
RBMT-1	N	0.60
UOW	Y	0.59
RBMT-2	N	0.57
KOC	Y	0.56
RBMT-5	N	0.54
CU-ZEMAN	Y	0.49
UPM	Y	0.34

English-French
868–1121 comparisons/system

System	C?	≥others
LIMSI ●⋆	Y+G	0.73
ONLINE-B ●	N	0.70
KIT ●⋆	Y+G	0.69
RWTH-HUCK	Y+G	0.65
LIUM	Y+G	0.64
RBMT-1	N	0.61
ONLINE-A	N	0.60
UEDIN	Y	0.58
RBMT-3	N	0.58
RBMT-5	N	0.55
UPPSALA	Y	0.55
JHU	Y	0.55
UPPSALA-FBK	Y	0.54
RBMT-4	N	0.49
RBMT-2	N	0.46
LATL-GENEVA	N	0.39
CU-ZEMAN	Y	0.20

German-English
741–998 comparisons/system

System	C?	≥others
ONLINE-B ●	N	0.72
CMU-DYER ●⋆	Y+G	0.66
ONLINE-A ●	N	0.66
RBMT-3	N	0.64
LINGUATEC	N	0.63
RBMT-4	N	0.61
RBMT-1	N	0.60
DFKI-XU	N	0.60
RWTH-WUEBKER ⋆	Y+G	0.59
KIT	Y+G	0.57
LIU	Y	0.57
LIMSI	Y+G	0.56
RBMT-5	N	0.56
UEDIN	Y	0.55
RBMT-2	N	0.54
CU-ZEMAN	Y	0.47
UPPSALA	Y	0.47
KOC	Y	0.45
JHU	Y+G	0.43
CST	Y	0.37

English-German
1051–1230 comparisons/system

System	C?	≥others
RBMT-3 ●	N	0.73
ONLINE-B ●	N	0.73
RBMT-1 ●	N	0.70
DFKI-FEDERMANN ●	N	0.68
DFKI-XU	N	0.67
RBMT-4 ●	N	0.66
RBMT-2 ●	N	0.66
ONLINE-A ●	N	0.65
LIMSI ⋆	Y+G	0.65
KIT ⋆	Y	0.64
UEDIN	Y	0.60
LIU	Y	0.59
RBMT-5	N	0.58
RWTH-FREITAG	Y	0.56
COPENHAGEN ⋆	Y	0.56
JHU	Y	0.54
KOC	Y	0.53
UOW	Y	0.53
CU-TAMCHYNA	Y	0.50
UPPSALA	Y	0.49
ILLC-UVA	Y	0.48
CU-ZEMAN	Y	0.38

C? indicates whether system is constrained: trained only using supplied training data, standard monolingual linguistic tools, and, optionally, LDC's English Gigaword. Eentries that used the Gigaword are marked with +G.

● indicates a **win**: no other system is statistically significantly better at p-level≤0.10 in pairwise comparison.

⋆ indicates a ***constrained* win**: no other *constrained* system is statistically better.

Table 8: Official results for the WMT11 translation task. Systems are ordered by their ≥others score, reflecting how often their translations won or tied pairwise comparisons. For detailed head-to-head comparisons, see Appendix A.

Czech-English
1036–1042 comparisons/combo

System	≥others
CMU-HEAFIELD-COMBO •	0.64
BBN-COMBO •	0.62
JHU-COMBO	0.58
UPV-PRHLT-COMBO	0.47

English-Czech
1788–1792 comparisons/combo

System	≥others
CMU-HEAFIELD-COMBO •	0.48
UPV-PRHLT-COMBO	0.41

German-English
811–927 comparisons/combo

System	≥others
CMU-HEAFIELD-COMBO •	0.70
RWTH-LEUSCH-COMBO	0.65
BBN-COMBO	0.61
UZH-COMBO •	0.60
JHU-COMBO	0.56
UPV-PRHLT-COMBO	0.52
QUAERO-COMBO	0.46
KOC-COMBO	0.45

English-German
1746–1752 comparisons/combo

System	≥others
CMU-HEAFIELD-COMBO •	0.61
UZH-COMBO •	0.58
UPV-PRHLT-COMBO	0.56
KOC-COMBO	0.46

Spanish-English
1132–1249 comparisons/combo

System	≥others
RWTH-LEUSCH-COMBO •	0.71
CMU-HEAFIELD-COMBO •	0.67
BBN-COMBO •	0.64
UPV-PRHLT-COMBO	0.64
JHU-COMBO	0.62
KOC-COMBO	0.56

English-Spanish
2360–2378 comparisons/combo

System	≥others
CMU-HEAFIELD-COMBO •	0.69
UOW-COMBO	0.63
UPV-PRHLT-COMBO	0.59
KOC-COMBO	0.58

French-English
820–916 comparisons/combo

System	≥others
BBN-COMBO •	0.67
RWTH-LEUSCH-COMBO •	0.63
CMU-HEAFIELD-COMBO	0.62
JHU-COMBO •	0.59
LIUM-COMBO	0.53
UPV-PRHLT-COMBO	0.53

English-French
586–587 comparisons/combo

System	≥others
CMU-HEAFIELD-COMBO •	0.51
UPV-PRHLT-COMBO	0.43

- indicates a **win**: no other system combination is statistically significantly better at p-level≤0.10 in pairwise comparison.

Table 9: Official results for the WMT11 system combination task. Systems are ordered by their ≥others score, reflecting how often their translations won or tied pairwise comparisons. For detailed head-to-head comparisons, see Appendix A.

Haitian Creole (Clean)-English
(individual systems)
1256–1435 comparisons/system

System	≥others
BM-I2R •	0.71
CMU-DENKOWSKI	0.66
CMU-HEWAVITHARANA	0.64
UMD-EIDELMAN	0.63
UPPSALA	0.57
LIU	0.55
UMD-HU	0.52
HYDERABAD	0.43
KOC	0.31

Haitian Creole (Raw)-English
(individual systems)
1065–1136 comparisons/system

System	≥others
BM-I2R •	0.65
CMU-HEWAVITHARANA	0.60
CMU-DENKOWSKI	0.59
LIU	0.55
UMD-EIDELMAN	0.52
JHU	0.41

Haitian Creole (Clean)-English
(system combinations)
896–898 comparisons/combo

System	≥others
CMU-HEAFIELD-COMBO •	0.52
UPV-PRHLT-COMBO	0.48
KOC-COMBO	0.38

Haitian Creole (Raw)-English
(system combinations)
600–600 comparisons/combo

System	≥others
CMU-HEAFIELD-COMBO	0.47
UPV-PRHLT-COMBO	0.43

- indicates a **win**: no other system is statistically significantly better at p-level≤0.10 in pairwise comparison.

Table 10: Official results for the WMT11 featured translation task (Haitian Creole SMS into English). Systems are ordered by their ≥others score, reflecting how often their translations won or tied pairwise comparisons. For detailed head-to-head comparisons, see Appendix A.

Tables 8–10 show the system ranking for each of the translation tasks. For each language pair, we define a system as 'winning' if no other system was found statistically significantly better (using the Sign Test, at $p \leq 0.10$). In some cases, multiple systems are listed as winners, either due to a large number of participants or a low number of judgments per system pair, both of which are factors that make it difficult to achieve statistical significance.

We start by examining the results for the individual system track for the European languages (Table 8). In Spanish↔English and German↔English, unconstrained systems are observed to perform better than constrained systems. In other language pairs, particularly French↔English, constrained systems are found to be able to be on the same level or outperform unconstrained systems. It also seems that making use of the Gigaword corpora is likely to yield better systems, even when translating out of English, as in English-French and English-German. For English-German the rule-based MT systems performed well.

Of the participating teams, there is no individual system clearly outperforming all other systems across the different language pairs. However, one of the crawled systems, ONLINE-B, performs consistently well, being one of the winners in all eight language pairs.

As for the system combination track (Table 9), the CMU-HEAFIELD-COMBO entry performed quite well, being a winner in seven out of eight language pairs. This performance is carried over to the Haitian Creole task, where it again comes out on top (Table 10). In the *individual* track of the Haitian Creole task, BM-I2R is the sole winner in both the 'clean' and 'raw' tracks.

6 Evaluation Task

In addition to allowing us to analyze the translation quality of different systems, the data gathered during the manual evaluation is useful for validating automatic evaluation metrics. Our evaluation shared task is similar to the MetricsMATR workshop (Metrics for MAchine TRanslation) that NIST runs (Przybocki et al., 2008; Callison-Burch et al., 2010). Table 11 lists the participants in this task, along with their metrics.

A total of 21 metrics and their variants were submitted to the evaluation task by 9 research groups. We asked metrics developers to score the outputs of the machine translation systems and system combinations at the system-level and at the segment-level. The system-level metrics scores are given in the Appendix in Tables 39–48. The main goal of the evaluation shared task is not to score the systems, but instead to validate the use of automatic metrics by measuring how strongly they correlate with human judgments. We used the human judgments collected during the manual evaluation for the translation task and the system combination task to calculate how well metrics correlate at system-level and at the segment-level.

This year the strongest metric was a new metric developed by Columbia and ETS called MTeRater-Plus. MTeRater-Plus is a machine-learning-based metric that use features from ETS's e-rater, an automated essay scoring engine designed to assess writing proficiency (Attali and Burstein, 2006). The features include sentence-level and document-level information. Some examples of the e-rater features include:

- Preposition features that calculate the probability of prepositions appearing in the given context of a sentence (Tetreault and Chodorow, 2008)
- Collocation features that indicate whether the collocations in the document are typical of native use (Futagi et al., 2008).
- A sentence fragment feature that counts the number of ill-formed sentences in a document.
- A feature that counts the number of words with inflection errors
- A feature that counts the the number of article errors in the sentence citeHan2006.

MTeRater uses only the e-rater features, and measures fluency without any need for reference translations. MTeRater-Plus is a meta-metric that incorporates adequacy by combining MTeRater with other MT evaluation metrics and heuristics that take the reference translations into account.

Please refer to the proceedings for papers providing detailed descriptions of all of the metrics.

Metric IDs	Participant
AMBER, AMBER-NL, AMBER-IT	National Research Council Canada (Chen and Kuhn, 2011)
F15, F15G3	Koç University (Bicici and Yuret, 2011)
METEOR-1.3-ADQ, METEOR-1.3-RANK	Carnegie Mellon University (Denkowski and Lavie, 2011a)
MTERATER, MTERATER-PLUS	Columbia / ETS (Parton et al., 2011)
MP4IBM1, MPF, WMPF	DFKI (Popović, 2011; Popović et al., 2011)
PARSECONF	DFKI (Avramidis et al., 2011)
ROSE, ROSE-POS	The University of Sheffield (Song and Cohn, 2011)
TESLA-B, TESLA-F, TESLA-M	National University of Singapore (Dahlmeier et al., 2011)
TINE	University of Wolverhampton (Rios et al., 2011)
BLEU	provided baseline (Papineni et al., 2002)
TER	provided baseline (Snover et al., 2006)

Table 11: Participants in the evaluation shared task. For comparison purposes, we include the BLEU and TER metrics as baselines.

	EN-CZ - 10 SYSTEMS	EN-DE - 22 SYSTEMS	EN-ES - 15 SYSTEMS	EN-FR - 17 SYSTEMS	AVERAGE	AVERAGE W/O CZ
System-level correlation for translation out of English						
TESLA-M		.90	**.95**	**.96**		**.94**
TESLA-B		.81	.90	.91		.87
MPF	.72	.63	.87	.89	**.78**	.80
WMPF	.72	.61	.87	.89	.77	.79
MP4IBM1	**-.76**	**-.91**	-.71	-.61	.75	.74
ROSE	.65	.41	.90	.86	.71	.73
BLEU	.65	.44	.87	.86	.70	.72
AMBER-TI	.56	.54	.88	.84	.70	.75
AMBER	.56	.53	.87	.84	.70	.74
AMBER-NL	.56	.45	.88	.83	.68	.72
F15G3	.50	.30	.89	.84	.63	.68
METEOR$_{rank}$	.65	.30	.74	.85	.63	.63
F15	.52	.19	.86	.85	.60	.63
TER	-.50	-.12	-.81	-.84	.57	.59
TESLA-F		.86	.80	-.83		.28

Table 12: System-level Spearman's rho correlation of the automatic evaluation metrics with the human judgments for translation out of English, ordered by average absolute value. We did not calculate correlations with the human judgments for the system combinations for the out of English direction, because none of them had more than 4 items.

6.1 System-Level Metric Analysis

We measured the correlation of the automatic metrics with the human judgments of translation quality at the system-level using Spearman's rank correlation coefficient ρ. We converted the raw scores assigned to each system into ranks. We assigned a human ranking to the systems based on the percent of time that their translations were judged to be better than or equal to the translations of any other system in the manual evaluation. The reference was not included as an extra translation.

When there are no ties, ρ can be calculated using the simplified equation:

$$\rho = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)}$$

where d_i is the difference between the rank for system$_i$ and n is the number of systems. The possible values of ρ range between 1 (where all systems are ranked in the same order) and -1 (where the systems are ranked in the reverse order). Thus an automatic evaluation metric with a higher absolute value for ρ is making predictions that are more similar to the human judgments than an automatic evaluation metric with a lower absolute ρ.

The system-level correlations are shown in Table 13 for translations into English, and Table 12 out of English, sorted by average correlation across the language pairs. The highest correlation for each language pair and the highest overall average are bolded. This year, nearly all of the metrics

	CZ-EN - 8 SYSTEMS	DE-EN - 20 SYSTEMS	DE-EN - 8 COMBOS	ES-EN - 15 SYSTEMS	ES-EN - 6 COMBOS	FR-EN - 18 SYSTEMS	FR-EN - 6 COMBOS	AVERAGE (EUROPEAN LANGS)	HT-EN (CLEAN) - 9 SYSTEMS	HT-EN (RAW) - 6 SYSTEMS	AVERAGE (ALL LANGS)
System-level correlation for metrics scoring translations into English											
MTERATER-PLUS	-.95	-.90	-.93	-.91	-.94	-.93	-.77	**.90**	-.82	-.54	.85
TINE-SRL-MATCH	.95	.69	.95	.95	**1.00**	.87	.66	.87			
TESLA-F	.95	.70	**.98**	**.96**	.94	.90	.60	.86	.93	.83	**.87**
TESLA-B	**.98**	.88	**.98**	.91	.94	.91	.31	.84	.93	.83	.85
MTERATER	-.91	-.88	-.91	-.88	-.89	-.79	-.60	.83	.13	.77	.55
METEOR-1.3-ADQ	.93	.68	.91	.91	.83	.93	.66	.83	**.95**	.77	.84
TESLA-M	.95	**.94**	.95	.82	.94	.87	.31	.83	.95	.83	.84
METEOR-1.3-RANK	.91	.71	.91	.88	.77	.93	.66	.82	.95	.83	.84
AMBER-NL	.88	.58	.91	.88	.94	.94	.60	.82			
AMBER-TI	.88	.63	.93	.85	.83	.94	.60	.81			
AMBER	.88	.59	.91	.86	.83	**.95**	.60	.80			
MPF	.95	.69	.91	.83	.60	.87	.54	.77	.95	.77	.79
WMPF	.95	.66	.86	.83	.60	.87	.54	.76	.93	.77	.78
F15	.93	.45	.88	.96	.49	.87	.60	.74			
F15G3	.93	.48	.83	.94	.49	.88	.60	.74			
ROSE	.88	.59	.83	.92	.60	.86	.26	.70	.93	.77	.74
BLEU	.88	.48	.83	.90	.49	.85	.43	.69	.90	.83	.73
TER	-.83	-.33	-.64	-.89	-.37	-.77	**-.89**	.67	-.93	-.83	.72
MP4IBM1	-.91	-.56	-.50	-.12	-.43	-.08	.14	.35			
DFKI-PARSECONF		.31	.52								

Table 13: System-level Spearman's rho correlation of the automatic evaluation metrics with the human judgments for translation into English, ordered by average absolute value for the European languages. We did not calculate correlations with the human judgments for the system combinations for Czech to English and for Haitian Creole to English, because they had too few items (≤ 4) for reliable statistics.

	FR-EN (6337 PAIRS)	DE-EN (8950 PAIRS)	ES-EN (5974 PAIRS)	CZ-EN (3695 PAIRS)	AVERAGE
Segment-level correlation for translations into English					
MTERATER-PLUS	**.30**	**.36**	**.45**	**.36**	**.37**
TESLA-F	.28	.24	.39	.32	.31
TESLA-B	.28	.26	.36	.29	.30
METEOR-1.3-RANK	.23	.25	.38	.28	.29
METEOR-1.3-ADQ	.24	.25	.37	.27	.28
MPF	.25	.23	.34	.28	.28
AMBER-TI	.24	.26	.33	.27	.28
AMBER	.24	.25	.33	.27	.27
WMPF	.24	.23	.34	.26	.27
AMBER-NL	.24	.24	.30	.27	.26
MTERATER	.19	.26	.33	.24	.26
TESLA-M	.21	.23	.29	.23	.24
TINE-SRL-MATCH	.20	.19	.30	.24	.23
F15G3	.17	.15	.29	.21	.21
F15	.16	.14	.27	.22	.20
MP4IBM1	.15	.16	.18	.12	.15
DFKI-PARSECONF	n/a	.24	n/a	n/a	

Table 14: Segment-level Kendall's tau correlation of the automatic evaluation metrics with the human judgments for translation into English, ordered by average correlation.

	EN-FR (6934 PAIRS)	EN-DE (10732 PAIRS)	EN-ES (8837 PAIRS)	EN-CZ (11651 PAIRS)	AVERAGE
Segment-level correlation for translations out of English					
AMBER-TI	**.32**	.22	**.31**	.21	**.27**
AMBER	.31	.21	**.31**	**.22**	.26
MPF	.31	**.22**	.30	.20	.26
WMPF	.31	**.22**	.29	.19	.25
AMBER-NL	.30	.19	.29	.20	.25
METEOR-1.3-RANK	.31	.14	.26	.19	.23
F15G3	.26	.08	.22	.13	.17
F15	.26	.07	.22	.12	.17
MP4IBM1	.21	.13	.13	.06	.13
TESLA-B	.29	.20	.28	n/a	
TESLA-M	.25	.18	.27	n/a	
TESLA-F	.30	.19	.26	n/a	

Table 15: Segment-level Kendall's tau correlation of the automatic evaluation metrics with the human judgments for translation out of English, ordered by average correlation.

had stronger correlation with human judgments than BLEU. The metrics that had the strongest correlation this year included two metrics, MTeRater and TINE, as well as metrics that have demonstrated strong correlation in previous years like TESLA and Meteor.

6.2 Segment-Level Metric Analysis

We measured the metrics' segment-level scores with the human rankings using Kendall's tau rank correlation coefficient. The reference was not included as an extra translation.

We calculated Kendall's tau as:

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

where a concordant pair is a pair of two translations of the same segment in which the ranks calculated from the same human ranking task and from the corresponding metric scores agree; in a discordant pair, they disagree. In order to account for accuracy- vs. error-based metrics correctly, counts of concordant vs. discordant pairs were calculated specific to these two metric types. The possible values of τ range between 1 (where all pairs are concordant) and -1 (where all pairs are discordant). Thus an automatic evaluation metric with a higher value for τ is making predictions that are more similar to the human judgments than an automatic evaluation metric with a lower τ.

We did not include cases where the human ranking was tied for two systems. As the metrics produce absolute scores, compared to five relative ranks in the human assessment, it would be potentially unfair to the metric to count a slightly different metric score as discordant with a tie in the relative human rankings. A tie in automatic metric rank for two translations was counted as discordant with two corresponding non-tied human judgments.

The correlations are shown in Table 14 for translations into English, and Table 15 out of English, sorted by average correlation across the four language pairs. The highest correlation for each language pair and the highest overall average are

ID	Participant	Metric Name
CMU-METEOR	Carnegie Mellon University	METEOR (Denkowski and Lavie, 2011a)
CU-SEMPOS-BLEU	Charles University	SemPOS/BLEU (Macháček and Bojar, 2011)
NUS-TESLA-F	National University of Singapore	TESLA-F (Dahlmeier et al., 2011)
RWTH-CDER	RWTH Aachen	CDER (Leusch and Ney, 2009)
SHEFFIELD-ROSE	The University of Sheffield	ROSE (single reference) (Song and Cohn, 2011)
STANFORD-DCP	Stanford	DCP (based on Liu and Gildea (2005))
BLEU	provided baseline	BLEU
BLEU-SINGLE	provided baseline	BLEU (single reference)

Table 16: Participants in the tunable-metric shared task. For comparison purposes, we included two BLEU-optimized systems in the evaluation as baselines.

bolded. There is a clear winner for the metrics that score translations into English: the MTeRater-Plus metric (Parton et al., 2011) has the highest segment level correlation across the board. For metrics that score translation into other languages, there is not such a clear-cut winner. The AMBER metric variants do well, as do MPF and WMPF.

7 Tunable Metrics Task

This year we introduced a new shared task that focuses on using evaluation metrics to tune the parameters of a statistical machine translation system. The intent of this task was to get researchers who develop automatic evaluation metrics for MT to work on the problem of using their metric to optimize the parameters of MT systems. Previous workshops have demonstrated that a number of metrics perform better than BLEU in terms of having stronger correlation with human judgments about the rankings of multiple machine translation systems. However, most MT system developers still optimize the parameters of their systems to BLEU. Here we aim to investigate the question of whether better metrics will result in better quality output when a system is optimized to them.

Because this was the first year that we ran the tunable metrics task, participation was limited to a few groups on an invitation-only basis. Table 16 lists the participants in this task. Metrics developers were invited to integrate their evaluation metric into a MERT optimization routine, which was then used to tune the parameters of a fixed statistical machine translation system. We evaluated whether the system tuned on their metrics produced higher-quality output than the baseline system that was tuned to BLEU, as is typically done. In order to evaluate whether the quality was better, we conducted a manual evaluation, in the same fashion that we evaluate the different MT systems submitted to the shared translation task.

We provide the participants with a fixed MT system for Urdu-English, along with a small parallel set to be used for tuning. Specifically, we provide developers with the following components:

- **Decoder** - the Joshua decoder was used in this pilot.
- **Decoder configuration file** - a Joshua configuration file that ensures all systems use the same search parameters.
- **Translation model** - an Urdu-to-English translation model, with syntax-based SCFG rules (Baker et al., 2010).
- **Language model** - a large 5-gram language model trained on the English Gigaword corpus
- **Development set** - a development set, with 4 English reference sets, to be used to optimize the system parameters.
- **Test set** - a test set consisting of 883 Urdu sentences, to be translated by the tuned system (no references provided).
- **Optimization routine** - we provide an implementation of minimum error rate training that allows new metrics to be easily integrated as the objective function.

Tunable Metrics Task
1324–1484 comparisons/system

System	≥**others**	>others
BLEU •	**0.79**	0.28
BLEU-SINGLE •	**0.77**	0.27
CMU-METEOR •	**0.76**	0.27
RWTH-CDER	**0.76**	0.26
CU-SEMPOS-BLEU •	**0.74**	0.29
STANFORD-DCP •	**0.73**	0.27
NUS-TESLA-F	**0.68**	0.28
SHEFFIELD-ROSE	**0.05**	0.00

- indicates a **win**: no other system combination is statistically significantly better at p-level≤0.10 in pairwise comparison.

Table 17: Official results for the WMT11 tunable-metric task. Systems are ordered by their ≥others score, reflecting how often their translations won or tied pairwise comparisons. The > column reflects how often a system strictly won a pairwise comparison.

	REF	BLEU	BLEU-SINGLE	CMU-METEOR	CU-SEMPOS-BLEU	NUS-TESLA-F	RWTH-CDER	SHEFFIELD-ROSE	STANFORD-DCP
REF	–	.15‡	.11‡	.13‡	.09‡	.09‡	.10‡	.00‡	.11‡
BLEU	**.78**‡	–	.15	**.11**	.20	.19†	.13⋆	.01‡	.14
BLEU-SINGLE	**.82**‡	**.20**	–	.11	.16	.21	.11	.00‡	**.20**
CMU-METEOR	**.84**‡	.09	**.15**	–	**.21**	.20	**.19**	.00‡	.19
CU-SEMPOS-BLEU	**.82**‡	**.23**	**.21**	.21	–	.12‡	.18	.00‡	.21
NUS-TESLA-F	**.80**‡	**.32**†	**.31**	**.28**	**.28**‡	–	**.31**	.00‡	.28
RWTH-CDER	**.79**‡	**.22**⋆	**.16**	.16	**.22**	.23	–	.00‡	.15
SHEFFIELD-ROSE	**.98**‡	**.93**‡	**.93**‡	**.96**‡	**.95**‡	**.95**‡	**.93**‡	–	**.94**‡
STANFORD-DCP	**.82**‡	**.17**	.18	**.26**	**.27**	.28	.15	.00‡	–
> others	.83	.28	.27	.27	**.29**	.28	.26	.00	.27
>= others	.90	**.79**	.77	.76	.74	.68	.76	.05	.73

Table 18: Head to head comparisons for the tunable metrics task. The numbers indicate how often the system in the column was judged to be better than the system in the row. The difference between 100 and the sum of the corresponding cells is the percent of time that the two systems were judged to be equal.

We provided the metrics developers with Omar Zaidan's Z-MERT software (Zaidan, 2009), which implements Och (2003)'s minimum error rate training procedure. Z-MERT is designed to be modular with respect to the objective function, and allows BLEU to be easily replaced with other automatic evaluation metrics. Metric developers incorporated their metrics into Z-MERT by subclassing the EvaluationMetric.java abstract class. They ran Z-MERT on the dev set with the provided decoder/models, and created a weight vector for the system parameters.

Each team produced a distinct final weight vector, which was used to produce English translations of sentences in the test set. The different translations produced by tuning the system to different metrics were then evaluated using the manual evaluation pipeline.[7]

7.1 Results of the Tunable Metrics Task

The results of the evaluation are in Table 18. The scores show that the entries were quite close to each other, with the notable exception of the SHEFFIELD-ROSE-tuned system, which produced overly-long and erroneous output (possibly due to an implementation issue). This is also evident from the fact that 38% of pairwise comparisons indicated a **tie** between the two systems, with the tie rate increasing to a full 47% when excluding comparisons involving the reference. This is a very high tie rate – the corresponding figure in, say, European language pairs (individual systems) is only 21%.

What makes the different entries appear even more closely-matched is that the ranking changes significantly when ordering systems by their >others score rather than the ≥others score (i.e. when rewarding only wins, and not rewarding ties). NUS-TESLA-F goes from being a bottom entry to being a top entry, with CU-SEMPOS-BLEU also benefiting, changing from the middle to the top rank.

Either way, we see that a BLEU -tuned system is performing just as well as systems tuned to the other metrics. This might be an indication that some work remains to be done before a move away from BLEU-tuning is fully justified. On the other hand, the close results might be an artifact of the language pair choice. Urdu-English translation is still a relatively difficult problem, and MT outputs are still of a relatively low quality. It might be the case that human annotators are simply not very good at distin-

[7] We also recased and detokenized each system's output, to ensure the outputs are more readable and easier to evaluate.

guishing one bad translation from another bad translation, especially at such a fine-grained level.

It is worth noting that the designers of the TESLA family replicated the setup of this tunable metric task for three European language pairs, and found that human judges *did* perceive a difference in quality between a TESLA-tuned system and a BLEU -tuned system (Liu et al., 2011).

7.2 Anticipated Changes Next Year

This year's effort was a pilot of the task, so we intentionally limited the task to some degree, to make it easier to iron out the details. Possible changes for next year include:

- More language pairs / translations into languages other than English. This year we focus on Urdu-English because the language pair requires a lot of reordering, and our syntactic model has more parameters to optimize than the standard Hiero and phrase-based models.
- Provide some human judgments about the model's output, so that people can experiment with regression models.
- Include a single reference track along with the multiple reference track. Some metrics may be better at dealing with the (more common) case of there being only a single reference translation available for every source sentence.
- Allow for experimentation with the MIRA optimization routine instead of MERT. MIRA can scale to a greater number of features, but requires that metrics be decomposable.

8 Summary

As in previous editions of this workshop we carried out an extensive manual and automatic evaluation of machine translation performance for translating from European languages into English, and vice versa.

The number of participants grew slightly compared to previous editions of the WMT workshop, with 36 groups from 27 institutions participating in the translation task of WMT11, 10 groups from 10 institutions participating in the system combination task, and 10 groups from 8 institutions participating in the featured translation task (Haitian Creole SMS into English).

This year was also the first time that we included a language pair (Haitian-English) with non-European source language and with very limited resources for the source language side. Also the genre of the Haitian-English task differed from previous WMT tasks as the Haitian-English translations are SMS messages.

WMT11 also introduced a new shared task focusing on evaluation metrics to tune the parameters of a statistical machine translation system in which 6 groups have participated.

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.[8]

Acknowledgments

This work was supported in parts by the EuroMatrixPlus project funded by the European Commission (7th Framework Programme), the GALE program of the US Defense Advanced Research Projects Agency, Contract No. HR0011-06-C-0022, the US National Science Foundation under grant IIS-0713448, and the CoSyne project FP7-ICT-4-248531 funded by the European Commission. The views and findings are the authors' alone. A big thank you to Ondřej Bojar, Simon Carter, Christian Federmann, Will Lewis, Rob Munro and Hervé Saint-Amand, and to the shared task participants.

References

Vera Aleksic and Gregor Thurmair. 2011. Personal Translator at WMT2011. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*.

Alexandre Allauzen, Hélène Bonneau-Maynard, Hai-Son Le, Aurélien Max, Guillaume Wisniewski, François Yvon, Gilles Adda, Josep Maria Crego, Adrien Lardilleux, Thomas Lavergne, and Artem Sokolov. 2011. LIMSI @ WMT11. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*.

Yigal Attali and Jill Burstein. 2006. Automated essay scoring with e-rater v.2.0. *Journal of Technology, Learning, and Assessment*, 4(3):159–174.

Eleftherios Avramidis, Maja Popović, David Vilar, and Aljoscha Burchardt. 2011. Evaluate with confidence

[8] http://statmt.org/wmt11/results.html

estimation: Machine ranking of translation outputs using grammatical features. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*.

Wilker Aziz, Miguel Rios, and Lucia Specia. 2011. Shallow semantic trees for SMT. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*.

Kathryn Baker, Michael Bloodgood, Chris Callison-Burch, Bonnie Dorr, Scott Miller, Christine Piatko, Nathaniel W. Filardo, and Lori Levin. 2010. Semantically-informed syntactic machine translation: A tree-grafting approach. In *Proceedings of AMTA*.

Loïc Barrault. 2011. MANY improvements for WMT'11. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*.

Ergun Bicici and Deniz Yuret. 2011. RegMT system for machine translation, system combination, and evaluation. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*.

Ondřej Bojar and Aleš Tamchyna. 2011. Improving translation model by monolingual data. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*.

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

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

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

Chris Callison-Burch, Philipp Koehn, Christof Monz, Kay Peterson, Mark Przybocki, and Omar F. Zaidan. 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.

Boxing Chen and Roland Kuhn. 2011. Amber: A modified bleu, enhanced ranking metric. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*.

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

Antonio M. Corbí-Bellot, Mikel L. Forcada, Sergio Ortiz-Rojas, Juan Antonio Pérez-Ortiz, Gema Ramírez-Sánchez, Felipe Sánchez-Martínez, Iñaki Alegria, Aingeru Mayor, and Kepa Sarasola. 2005. An open-source shallow-transfer machine translation engine for the romance languages of Spain. In *Proceedings of the European Association for Machine Translation*, pages 79–86.

Marta R. Costa-jussà and Rafael E. Banchs. 2011. The BM-I2R Haitian-Créole-to-English translation system description for the WMT 2011 evaluation campaign. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*.

Daniel Dahlmeier, Chang Liu, and Hwee Tou Ng. 2011. TESLA at WMT 2011: Translation evaluation and tunable metric. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*.

Michael Denkowski and Alon Lavie. 2011a. Meteor 1.3: Automatic metric for reliable optimization and evaluation of machine translation systems. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*.

Michael Denkowski and Alon Lavie. 2011b. METEOR-Tuned Phrase-Based SMT: CMU French-English and Haitian-English Systems for WMT 2011. Technical Report CMU-LTI-11-011, Language Technologies Institute, Carnegie Mellon University.

Chris Dyer, Kevin Gimpel, Jonathan H. Clark, and Noah A. Smith. 2011. The CMU-ARK German-English translation system. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*.

Vladimir Eidelman, Kristy Hollingshead, and Philip Resnik. 2011. Noisy SMS machine translation in low-density languages. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*.

Christian Federmann and Sabine Hunsicker. 2011. Stochastic parse tree selection for an existing RBMT system. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*.

Robert Frederking, Alexander Rudnicky, and Christopher Hogan. 1997. Interactive speech translation in the DIPLOMAT project. In *Proceedings of the ACL-1997 Workshop on Spoken Language Translation*.

Markus Freitag, Gregor Leusch, Joern Wuebker, Stephan Peitz, Hermann Ney, Teresa Herrmann, Jan Niehues, Alex Waibel, Alexandre Allauzen, Gilles Adda, Josep Maria Crego, Bianka Buschbeck, Tonio Wandmacher, and Jean Senellart. 2011. Joint WMT submission of the QUAERO project. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*.

Yoko Futagi, Paul Deane, Martin Chodorow, and Joel Tetreault. 2008. A computational approach to detecting collocation errors in the writing of non-native speakers of English. *Computer Assisted Language Learning Journal*.

Jesús González-Rubio and Francisco Casacuberta. 2011. The UPV-PRHLT combination system for WMT 2011.

In *Proceedings of the Sixth Workshop on Statistical Machine Translation*.

Greg Hanneman and Alon Lavie. 2011. CMU syntax-based machine translation at WMT 2011. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*.

Christian Hardmeier, Jörg Tiedemann, Markus Saers, Marcello Federico, and Mathur Prashant. 2011. The Uppsala-FBK systems at WMT 2011. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*.

Kenneth Heafield and Alon Lavie. 2011. CMU system combination in WMT 2011. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*.

Teresa Herrmann, Mohammed Mediani, Jan Niehues, and Alex Waibel. 2011. The Karlsruhe Institute of Technology translation systems for the WMT 2011. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*.

Sanjika Hewavitharana, Nguyen Bach, Qin Gao, Vamshi Ambati, and Stephan Vogel. 2011. CMU Haitian Creole-English translation system for WMT 2011. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*.

Maria Holmqvist, Sara Stymne, and Lars Ahrenberg. 2011. Experiments with word alignment, normalization and clause reordering for SMT between English and German. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*.

Chang Hu, Philip Resnik, Yakov Kronrod, Vladimir Eidelman, Olivia Buzek, and Benjamin B. Bederson. 2011. The value of monolingual crowdsourcing in a real-world translation scenario: Simulation using Haitian Creole emergency SMS messages. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*.

Matthias Huck, Joern Wuebker, Christoph Schmidt, Markus Freitag, Stephan Peitz, Daniel Stein, Arnaud Dagnelies, Saab Mansour, Gregor Leusch, and Hermann Ney. 2011. The RWTH Aachen machine translation system for WMT 2011. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*.

Maxim Khalilov and Khalil Sima'an. 2011. ILLC-UvA translation system for EMNLP-WMT 2011. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*.

Philipp Koehn and Christof Monz. 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.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the ACL-2007 Demo and Poster Sessions*, Prague, Czech Republic.

Oliver Lacey-Hall. 2011. The guardian's poverty matters blog: How remote teams can help the rapid response to disasters, March.

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

Gregor Leusch and Hermann Ney. 2009. Edit distances with block movements and error rate confidence estimates. *Machine Translation*, 23:129–140.

Gregor Leusch, Markus Freitag, and Hermann Ney. 2011. The RWTH system combination system for WMT 2011. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*.

William Lewis, Robert Munro, and Stephan Vogel. 2011. Crisis MT: Developing a cookbook for MT in crisis situations. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*.

William D. Lewis. 2010. Haitian Creole: How to build and ship an MT engine from scratch in 4 days, 17hours, & 30 minutes. In *Proceedings of EAMT 2010*.

Zhifei Li, Chris Callison-Burch, Chris Dyer, Juri Ganitkevitch, Ann Irvine, Sanjeev Khudanpur, Lane Schwartz, Wren Thornton, Ziyuan Wang, Jonathan Weese, and Omar Zaidan. 2010. Joshua 2.0: A toolkit for parsing-based machine translation with syntax, semirings, discriminative training and other goodies. In *Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR*, Uppsala, Sweden, July.

Ding Liu and Daniel Gildea. 2005. Syntactic features for evaluation of machine translation. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 25–32.

Chang Liu, Daniel Dahlmeier, and Hwee Tou Ng. 2011. Better evaluation metrics lead to better machine translation. In *Proceedings of EMNLP*.

Verónica López-Ludeña and Rubén San-Segundo. 2011. UPM system for the translation task. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*.

Matouš Macháček and Ondřej Bojar. 2011. Approximating a deep-syntactic metric for MT evaluation and tuning. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*.

David Mareček, Rudolf Rosa, Petra Galuščáková, and Ondřej Bojar. 2011. Two-step translation with grammatical post-processing. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*.

Robert Munro. 2010. Crowdsourced translation for emergency response in Haiti: the global collaboration of local knowledge. In *Proceedings of the AMTA Workshop on Collaborative Crowdsourcing for Translation*.

Douglas W. Oard and Franz Josef Och. 2003. Rapid-response machine translation for unexpected languages. In *Proceedings of MT Summit IX*.

Douglas W. Oard. 2003. The surprise language exercises. *ACM Transactions on Asian Language Information Processing*, 2(2):79–84.

Franz Josef Och. 2003. Minimum error rate training for statistical machine translation. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics (ACL-2003)*, Sapporo, Japan.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL-2002)*, Philadelphia, Pennsylvania.

Kristen Parton, Joel Tetreault, Nitin Madnani, and Martin Chodorow. 2011. E-rating machine translation. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*.

Martin Popel, David Mareček, Nathan Green, and Zdeněk Žabokrtský. 2011. Influence of parser choice on dependency-based MT. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*.

Maja Popović, David Vilar, Eleftherios Avramidis, and Aljoscha Burchardt. 2011. Evaluation without references: IBM1 scores as evaluation metrics. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*.

Maja Popović. 2011. Morphemes and POS tags for n-gram based evaluation metrics. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*.

Marion Potet, Raphaël Rubino, Benjamin Lecouteux, Stéphane Huet, Laurent Besacier, Hervé Blanchon, and Fabrice Lefèvre. 2011. The LIGA (LIG/LIA) machine translation system for WMT 2011. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*.

Mark Przybocki, Kay Peterson, and Sebastian Bronsart. 2008. Official results of the NIST 2008 "Metrics for MAchine TRanslation" challenge (MetricsMATR08). In *AMTA-2008 workshop on Metrics for Machine Translation*, Honolulu, Hawaii.

Miguel Rios, Wilker Aziz, and Lucia Specia. 2011. TINE: A metric to assess MT adequacy. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*.

Christian Rishøj and Anders Søgaard. 2011. Factored translation with unsupervised word clusters. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*.

Antti-Veikko Rosti, Bing Zhang, Spyros Matsoukas, and Richard Schwartz. 2011. Expected BLEU training for graphs: BBN system description for WMT11 system combination task. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*.

Víctor M. Sánchez-Cartagena, Felipe Sánchez-Martínez, and Juan Antonio Pérez-Ortiz. 2011. The Universitat d'Alacant hybrid machine translation system for WMT 2011. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*.

Holger Schwenk, Patrik Lambert, Loïc Barrault, Christophe Servan, Sadaf Abdul-Rauf, Haithem Afli, and Kashif Shah. 2011. LIUM's SMT machine translation systems for WMT 2011. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*.

Rico Sennrich. 2011. The UZH system combination system for WMT 2011. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*.

Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 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.

Xingyi Song and Trevor Cohn. 2011. Regression and ranking based optimisation for sentence level MT evaluation. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*.

Lucia Specia, Dhwaj Raj, and Marco Turchi. 2010. Machine translation evaluation versus quality estimation. *Machine Translation*, 24(1):39–50.

Sara Stymne. 2011. Spell checking techniques for replacement of unknown words and data cleaning for Haitian Creole SMS translation. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*.

Joel Tetreault and Martin Chodorow. 2008. The ups and downs of preposition error detection. In *Proceedings of COLING*, Manchester, UK.

Jonathan Weese, Juri Ganitkevitch, Chris Callison-Burch, Matt Post, and Adam Lopez. 2011. Joshua 3.0: Syntax-based machine translation with the Thrax grammar extractor. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*.

Eric Wehrli, Luka Nerima, and Yves Scherrer. 2009. Deep linguistic multilingual translation and bilingual dictionaries. In *Proceedings of the Fourth Workshop on Statistical Machine Translation*, pages 90–94.

Daguang Xu, Yuan Cao, and Damianos Karakos. 2011a. Description of the JHU system combination scheme for WMT 2011. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*.

Jia Xu, Hans Uszkoreit, Casey Kennington, David Vilar, and Xiaojun Zhang. 2011b. DFKI hybrid machine translation system for WMT 2011 - on the integration of SMT and RBMT. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*.

Omar F. Zaidan. 2009. Z-MERT: A fully configurable open source tool for minimum error rate training of machine translation systems. *The Prague Bulletin of Mathematical Linguistics*, 91:79–88.

Francisco Zamora-Martinez and Maria Jose Castro-Bleda. 2011. CEU-UPV English-Spanish system for WMT11. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*.

Daniel Zeman. 2011. Hierarchical phrase-based MT at the Charles University for the WMT 2011 shared task. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*.

A Pairwise System Comparisons by Human Judges

Tables 19–38 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. Bolding indicates the winner of the two systems. The difference between 100 and the sum of the complementary cells is the percent of time that the two systems were judged to be equal.

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 $\star$ 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.

B Automatic Scores

Tables 39–48 give the automatic scores for each of the systems.

C Meta-evaluation

Tables 49 and 50 give a detailed breakdown of intra- and inter-annotator agreement rates for all of manual evaluation tracks of WMT11, broken down by language pair.

	REF	CST	CU-BOJAR	CU-ZEMAN	JHU	ONLINE-B	SYSTRAN	UEDIN	UPPSALA
REF	–	.02‡	.04‡	.01‡	.04‡	.04‡	.04‡	.05‡	.04‡
CST	**.88**‡	–	**.49**‡	**.36**	**.49**†	**.59**‡	**.41**	**.58**‡	**.44**†
CU-BOJAR	**.91**‡	.27‡	–	.27‡	.30	**.48**‡	.28‡	**.41**†	**.41**
CU-ZEMAN	**.94**‡	.31	**.49**‡	–	**.47**‡	**.67**‡	**.47**†	**.64**‡	**.49**‡
JHU	**.89**‡	.29†	**.39**	.28‡	–	**.47**‡	.36	**.41**†	.36
ONLINE-B	**.84**‡	.20‡	.27‡	.19‡	.28‡	–	.24‡	.30	.27‡
SYSTRAN	**.91**‡	.31	**.49**‡	.30†	**.39**	**.59**‡	–	**.56**‡	.37
UEDIN	**.89**‡	.16‡	.25†	.16‡	.27†	**.36**	.23‡	–	.25†
UPPSALA	**.84**‡	.28†	.40	.24‡	**.37**	**.49**‡	**.38**	**.45**†	–
> others	.89	.23	.36	.23	.33	**.46**	.31	.43	.33
>= others	.96	.47	.60	.44	.57	.68	.51	**.69**	.57

Table 19: Ranking scores for entries in the Czech-English task (individual system track).

	REF	COMMERCIAL-1	COMMERCIAL-2	CU-BOJAR	CU-MARECEK	CU-POPEL	CU-TAMCHYNA	CU-ZEMAN	JHU	ONLINE-B	UEDIN
REF	–	.05‡	.04‡	.04‡	.04‡	.05‡	.05‡	.04‡	.03‡	.04‡	.04‡
COMMERCIAL-1	**.91**‡	–	.36	**.53**‡	**.50**‡	**.47**‡	**.44**⋆	.33‡	.33†	**.55**‡	**.45**†
COMMERCIAL-2	**.87**‡	**.42**	–	**.52**‡	**.47**⋆	**.47**‡	**.50**‡	.30‡	.40	**.50**‡	**.43**
CU-BOJAR	**.89**‡	.31‡	.31‡	–	**.29**	.41	.21†	.19‡	.27‡	**.42**⋆	.31⋆
CU-MARECEK	**.88**‡	.31‡	.37⋆	.27	–	.35†	.28	.21‡	.30‡	.39	.28†
CU-POPEL	**.85**‡	.33‡	.29‡	**.43**	**.45**†	–	**.41**	.27‡	.31‡	**.50**‡	.39
CU-TAMCHYNA	**.87**‡	.34⋆	.35‡	**.30**†	**.32**	.40	–	.22‡	.25‡	**.45**‡	.32
CU-ZEMAN	**.91**‡	**.47**‡	**.52**‡	**.56**‡	**.56**‡	**.55**‡	**.55**‡	–	**.44**‡	**.64**‡	**.54**‡
JHU	**.91**‡	**.43**†	**.41**	**.50**‡	**.47**‡	**.51**‡	**.51**‡	.31‡	–	**.52**‡	**.48**‡
ONLINE-B	**.86**‡	.27‡	.32‡	.33⋆	**.39**	.33‡	.29‡	.18‡	.23‡	–	.31‡
UEDIN	**.85**‡	.34†	.40	**.40**⋆	**.37**†	**.42**	**.36**	.24‡	.25‡	**.44**‡	–
> others	.88	.33	.34	.39	.39	.40	.36	.23	.28	**.44**	.35
>= others	.96	.51	.51	.64	.63	.58	.62	.43	.49	**.65**	.59

Table 20: Ranking scores for entries in the English-Czech task (individual system track).

	REF	CMU-DYER	CST	CU-ZEMAN	DFKI-XU	JHU	KIT	KOC	LIMSI	LINGUATEC	LIU	ONLINE-A	ONLINE-B	RBMT-1	RBMT-2	RBMT-3	RBMT-4	RBMT-5	RWTH-WUEBKER	UEDIN	UPPSALA
REF	–	.05‡	.02‡	.03‡	.04‡	.00‡	.08‡	.04‡	.00‡	.07‡	.05‡	.07‡	.14‡	.02‡	.08‡	.00‡	.06‡	.08‡	.02‡	.10‡	.08‡
CMU-DYER	**.95**‡	–	.18‡	.17‡	.33	.26⋆	.22‡	.12‡	.29⋆	.43	.23⋆	.43	**.54**	.32	.20†	.40	**.43**	**.48**	.31	.19†	.18‡
CST	**.96**‡	**.74**‡	–	**.42**	**.62**‡	**.35**	**.68**‡	**.44**‡	**.47**⋆	**.78**‡	**.62**‡	**.77**‡	**.73**‡	**.81**‡	**.70**‡	**.74**‡	**.67**‡	**.53**⋆	**.65**‡	**.47**	**.51**
CU-ZEMAN	**.97**‡	**.67**‡	.22	–	**.56**†	.26†	**.41**	.22⋆	**.48**	**.66**‡	**.46**	**.60**‡	**.62**‡	**.73**‡	**.57**†	**.60**†	**.62**‡	**.53**⋆	**.40**	**.44**	**.48**
DFKI-XU	**.94**‡	**.44**	.06‡	.24†	–	.10‡	.26	.17‡	**.49**†	**.47**	.21⋆	.42	**.45**	**.52**	.42	**.45**	**.51**	.39	**.40**	**.48**	.29
JHU	**1.00**‡	**.61**⋆	.33	**.55**†	**.64**‡	–	**.59**†	**.45**	**.51**⋆	**.59**	**.52**⋆	**.68**‡	**.63**†	**.62**‡	**.64**†	**.65**‡	**.58**†	**.46**	**.61**‡	**.44**	.38
KIT	**.87**‡	**.65**‡	.12‡	.21	**.44**	.23†	–	.34	**.40**	**.54**	.30	**.43**	**.57**†	.44	.43	.47	**.50**	**.53**	**.40**	.28	.17‡
KOC	**.96**‡	**.64**‡	.09‡	**.49**⋆	**.66**‡	.36	**.43**	–	**.43**	**.69**‡	**.57**†	**.69**‡	**.63**‡	**.62**†	.41	**.63**‡	**.59**	**.52**⋆	**.51**	**.59**⋆	**.40**
LIMSI	**.96**‡	**.54**⋆	.24⋆	.30	.22†	.25⋆	.38	.27	–	**.63**†	**.52**	**.43**	**.55**†	.43	**.43**	**.59**†	**.47**	.40	.41	.32	**.44**
LINGUATEC	**.91**‡	**.45**	.13‡	.24‡	.38	.32	.34	.18‡	.27†	–	.26†	**.45**	**.62**†	**.46**	.20‡	**.49**	**.53**	.36	.41	.32⋆	.29†
LIU	**.89**‡	**.49**⋆	.14‡	.29	**.54**⋆	.25⋆	**.48**	.24†	.31	**.64**†	–	**.47**	**.61**†	**.52**	**.46**	**.48**	**.50**	.23‡	**.48**	.37	.36
ONLINE-A	**.88**‡	**.47**	.12‡	.25‡	.42	.18‡	.41	.19‡	.39	.39	.30	–	.32	.26†	.28	.46	.36	.35	.42	.19‡	.27‡
ONLINE-B	**.78**‡	.38	.16‡	.23‡	.33	.28†	.26†	.16‡	.26†	.29†	.22†	**.38**	–	.23‡	.23†	.29⋆	.29⋆	.22‡	.27	.22†	.18‡
RBMT-1	**.96**‡	**.42**	.09‡	.18‡	.35	.21‡	**.51**	.23†	.43	.41	.38	**.56**†	**.62**‡	–	.31	**.46**	**.39**	.13	**.48**	**.50**	.30⋆
RBMT-2	**.86**‡	**.54**†	.15‡	.28†	**.48**	.29†	.43	.41	.39	**.55**‡	.44	**.51**	**.64**†	**.43**	–	**.55**⋆	**.47**	**.54**⋆	**.44**	**.41**	.29⋆
RBMT-3	**.92**‡	**.42**	.11‡	.27†	.32	.23‡	.47	.18‡	.19†	.34	.38	**.49**	**.55**⋆	.38	.26⋆	–	.36	.29⋆	.34	.33	.28†
RBMT-4	**.88**‡	.36	.19‡	.24‡	.38	.29†	.43	.38	.45	.32	.37	**.44**	**.56**⋆	.33	.34	**.45**	–	.35	.29⋆	**.51**	.24†
RBMT-5	**.92**‡	.45	.27⋆	.27⋆	**.45**	.32	.37	.27⋆	**.47**	**.47**	**.61**‡	**.55**	**.67**‡	**.26**	.24⋆	**.53**⋆	**.46**	–	.45	**.47**	.39
RWTH-WUEBKER	**.93**‡	**.50**	.23‡	.26	.33	.20‡	.24	.36	.41	**.44**	.39	**.47**	**.55**	.44	.38	**.53**	**.56**⋆	.45	–	.21	**.39**
UEDIN	**.88**‡	**.59**†	.24	.28	.28	.33	**.50**	.24⋆	**.45**	**.65**⋆	**.40**	**.67**‡	**.62**†	.34	.39	**.52**	.41	.36	**.43**	–	**.48**
UPPSALA	**.92**‡	**.64**‡	.27	.29	**.39**	**.44**	**.58**‡	.32	.41	**.66**†	**.53**	**.68**‡	**.69**‡	**.59**⋆	**.59**⋆	**.58**†	**.61**†	**.54**	.36	.31	–
> others	.92	.50	.17	.28	.40	.26	.40	.26	.38	.51	.40	.51	**.57**	.43	.38	.49	.47	.39	.41	.36	.32
>= others	.95	.66	.37	.47	.60	.43	.57	.45	.56	.63	.57	.66	**.72**	.60	.54	.64	.61	.56	.59	.55	.47

Table 21: Ranking scores for entries in the German-English task (individual system track).

	REF	COPENHAGEN	CU-TAMCHYNA	CU-ZEMAN	DFKI-FEDERMANN	DFKI-XU	ILLC-UVA	JHU	KIT	KOC	LIMSI	LIU	ONLINE-A	ONLINE-B	RBMT-1	RBMT-2	RBMT-3	RBMT-4	RBMT-5	RWTH-FREITAG	UEDIN	UOW	UPPSALA
REF	–	.08‡	.06‡	.00‡	.13‡	.02‡	.05‡	.05‡	.02‡	.02‡	.16‡	.06‡	.11‡	.07‡	.14‡	.14‡	.19‡	.11‡	.11‡	.16‡	.07‡	.07‡	.08‡
COPENHAGEN	**.85**‡	–	.31	.09‡	**.60**‡	**.39**	.25	.32	**.41**	.27	**.36**	**.34**	**.49**†	**.61**‡	**.56**‡	**.61**‡	**.64**‡	**.64**‡	**.60**	.26	**.49**	.30	.16
CU-TAMCHYNA	**.92**‡	**.37**	–	.13†	**.61**‡	**.48**†	.30	**.38**	**.58**†	.33	**.39**	**.41**⋆	**.55**†	**.57**‡	**.72**‡	**.69**‡	**.81**‡	**.49**	**.59**†	**.47**	**.39**	**.40**	**.43**
CU-ZEMAN	**1.00**‡	**.60**‡	**.41**†	–	**.76**‡	**.78**‡	**.51**†	**.47**⋆	**.64**‡	**.53**‡	**.66**‡	**.49**⋆	**.77**‡	**.68**‡	**.69**‡	**.64**‡	**.70**‡	**.64**‡	**.72**‡	**.55**‡	**.47**	**.44**	**.50**
DFKI-FEDERMANN	**.72**‡	.19‡	.17‡	.16‡	–	**.39**	.25‡	.38	.38	.24‡	.32	.29	.35	.40	**.43**	.33	**.39**	.19	.33⋆	.22†	.31	.11‡	.30
DFKI-XU	**.84**‡	.31	.21†	.08‡	.37	–	.25†	.32	.34	.12‡	**.37**	.30	.35	**.47**	**.54**⋆	.30	**.51**⋆	**.43**	.37	.20†	.22†	.25†	.14‡
ILLC-UVA	**.90**‡	**.39**	**.37**	.25†	**.63**‡	**.50**†	–	**.41**⋆	**.58**‡	.35	**.56**‡	**.38**	**.55**†	**.63**‡	**.61**‡	**.63**‡	**.71**‡	**.75**‡	**.62**†	**.33**	**.56**‡	**.38**	**.41**
JHU	**.91**‡	**.45**	.27	.27⋆	**.41**	**.40**	.20⋆	–	.37	.27	**.43**	**.50**‡	**.58**†	**.59**‡	**.43**	**.55**†	**.72**‡	**.50**	**.50**	**.50**⋆	**.47**	**.46**	.22†
KIT	**.87**‡	.24	.23†	.17‡	**.41**	**.43**	.26‡	.37	–	.16‡	**.51**	.27†	.37	**.45**⋆	**.47**	.39	**.58**†	**.53**	**.47**	.23‡	.24	.21‡	.17‡
KOC	**.95**‡	**.35**	**.35**	.13‡	**.61**‡	**.65**‡	**.38**	**.42**	**.57**‡	–	**.47**⋆	.33	**.47**⋆	**.62**‡	**.61**†	**.53**†	**.64**‡	**.63**‡	**.45**	.20	**.38**	**.37**	.18†
LIMSI	**.77**‡	.31	.26	.11‡	**.48**	.35	.18‡	.30	.33	.23⋆	–	.36	**.39**	**.50**†	**.52**	**.47**	**.48**	.39	.42	.18‡	.22†	.28	.14‡
LIU	**.84**‡	.32	.20⋆	.25⋆	**.51**	**.38**	.26	.21‡	**.51**†	**.35**	**.39**	–	**.51**	**.49**⋆	**.63**†	**.52**⋆	**.56**	**.48**⋆	**.56**	.29	**.38**	.25	.25
ONLINE-A	**.75**‡	.21†	.24†	.09‡	**.48**	**.41**	.22†	.30†	.37	.25⋆	.37	.37	–	**.46**	.37	**.41**	**.47**	.33	.44	.27⋆	.28	.22‡	.16‡
ONLINE-B	**.91**‡	.17‡	.15‡	.13‡	**.44**	.22	.17‡	.16‡	.20⋆	.15‡	.24†	.25⋆	.27	–	**.43**	.35	**.48**	.33	.17‡	.17‡	.26	.12‡	.20‡
RBMT-1	**.80**‡	.23‡	.11‡	.20‡	.37	.28⋆	.18‡	.29	.38	.25†	.36	.30†	**.41**	.38	–	.34	**.45**	.36	.02‡	.17‡	.17‡	.28⋆	.24†
RBMT-2	**.80**‡	.20‡	.10‡	.16‡	**.43**	**.38**	.20‡	.27†	**.45**	.22†	.36	.30⋆	.38	**.51**	**.43**	–	**.48**	**.40**	**.42**	.31⋆	.28⋆	.16‡	.25‡
RBMT-3	**.65**‡	.18‡	.14‡	.15‡	.37	.29⋆	.17‡	.22‡	.25†	.20‡	.27	.33	.33	.29	.30	.31	–	.34	.16‡	.24†	.35	.20‡	.11‡
RBMT-4	**.80**‡	.21‡	.28	.22‡	.19	.26	.09‡	.32	.29	.27‡	.39	.27⋆	**.43**	**.44**	**.38**	.38	**.45**	–	**.42**	.29⋆	.36	.27‡	.31⋆
RBMT-5	**.88**‡	.35	.31†	.15‡	**.54**⋆	**.51**	.26†	.34	.36	.36	**.44**	.35	.44	**.59**‡	**.37**‡	.33	**.62**‡	.38	–	.29	**.45**	.38	.30
RWTH-FREITAG	**.80**‡	**.31**	.27	.17‡	**.62**†	**.55**†	.19	.25⋆	**.56**‡	**.30**	**.49**‡	**.41**	**.53**⋆	**.59**‡	**.56**‡	**.53**⋆	**.62**†	**.57**⋆	**.45**	–	**.36**	**.38**	.24
UEDIN	**.82**‡	.27	.27	.27	**.46**	**.47**†	.17‡	.28	**.36**	.33	**.48**†	.27	**.47**	**.43**	**.75**‡	**.55**⋆	**.52**	**.50**	.43	.21	–	.35	.27
UOW	**.86**‡	**.39**	.21	.23	**.74**‡	**.53**†	.36	.38	**.64**‡	.20	**.38**	**.41**	**.74**‡	**.61**‡	**.56**⋆	**.64**‡	**.57**‡	**.65**‡	.38	.26	**.41**	–	.31
UPPSALA	**.79**‡	**.32**	.35	.29	**.54**	**.57**‡	.34	**.51**†	**.51**‡	**.45**†	**.53**‡	**.43**	**.73**‡	**.70**‡	**.55**†	**.64**‡	**.77**‡	**.57**⋆	**.55**	**.43**	**.33**	**.41**	–
> others	.84	.29	.24	.17	.48	.42	.24	.31	.42	.27	.40	.34	.46	.51	.51	.47	**.56**	.46	.41	.29	.34	.29	.25
>= others	.91	.56	.50	.38	.68	.67	.48	.54	.64	.53	.65	.59	.65	**.730**	.70	.66	**.732**	.66	.58	.56	.60	.53	.49

Table 22: Ranking scores for entries in the English-German task (individual system track).

	REF	ALACANT	CU-ZEMAN	HYDERABAD	KOC	ONLINE-A	ONLINE-B	RBMT-1	RBMT-2	RBMT-3	RBMT-4	RBMT-5	SYSTRAN	UEDIN	UFAL-UM	UPM
REF	–	.03‡	.02‡	.00‡	.02‡	.03‡	.12‡	.15‡	.04‡	.07‡	.05‡	.02‡	.03‡	.03‡	.03‡	.07‡
ALACANT	**.86**‡	–	.07‡	.08‡	.30	**.52**	.31	.27⋆	.29⋆	**.54**	**.49**	.32⋆	**.51**	.27†	.26†	.26⋆
CU-ZEMAN	**.98**‡	**.89**‡	–	**.48**	**.84**‡	**.85**‡	**.94**‡	**.90**‡	**.83**‡	**.87**‡	**.85**‡	**.78**‡	**.97**‡	**.79**‡	**.79**‡	**.91**‡
HYDERABAD	**.98**‡	**.86**‡	.27	–	**.88**‡	**.95**‡	**.92**‡	**.85**‡	**.96**‡	**.74**‡	**.82**‡	**.80**‡	**.88**‡	**.91**‡	**.80**‡	**.86**‡
KOC	**.93**‡	**.48**	.06‡	.06‡	–	.28	.39	.40	.34	.44	.38	.26†	**.59**†	.22‡	.20‡	.18‡
ONLINE-A	**.90**‡	.28	.02‡	.02‡	**.48**	–	**.32**	.34	.34	.26⋆	.34	.19‡	.35	.20‡	.11‡	.20‡
ONLINE-B	**.79**‡	**.33**	.04‡	.00‡	**.47**	.30	–	.24†	.31⋆	.31⋆	.27†	.25‡	.33	.27†	.21‡	.07‡
RBMT-1	**.81**‡	**.52**⋆	.05‡	.11‡	**.50**	**.57**	**.62**†	–	**.50**	.36	.34	.17	.40	.39	.34	.30⋆
RBMT-2	**.96**‡	**.61**⋆	.09‡	.04‡	**.52**	**.47**	**.59**⋆	.37	–	**.39**	**.46**	.27	**.58**†	.29†	.24†	.45
RBMT-3	**.88**‡	.31	.09‡	.13‡	.44	**.56**⋆	**.60**⋆	**.53**	.37	–	**.47**	.14‡	**.52**	.40	.23†	.31
RBMT-4	**.90**‡	.38	.08‡	.16‡	**.50**	**.53**	**.60**†	**.41**	.43	.38	–	**.43**	**.52**	.33⋆	.18‡	.22‡
RBMT-5	**.94**‡	**.61**⋆	.06‡	.10‡	**.54**†	**.70**‡	**.63**‡	**.37**	**.45**	**.59**‡	.41	–	**.66**‡	.42	**.50**	**.43**
SYSTRAN	**.92**‡	.33	.02‡	.10‡	.25†	**.53**	**.53**	**.42**	.30†	.36	.38	.27‡	–	.21‡	.41	.24‡
UEDIN	**.95**‡	**.63**†	.13‡	.02‡	**.63**‡	**.67**‡	**.59**†	**.47**	**.61**†	**.53**	**.59**⋆	.42	**.53**‡	–	.32†	**.45**
UFAL-UM	**.94**‡	**.63**†	.10‡	.11‡	**.56**‡	**.70**‡	**.74**‡	**.51**	**.61**†	**.59**†	**.74**‡	.36	**.47**	**.61**†	–	**.44**
UPM	**.85**‡	**.54**⋆	.02‡	.03‡	**.62**‡	**.61**‡	**.81**‡	**.59**⋆	.45	**.55**	**.68**‡	.40	**.60**‡	.42	.38	–
> others	.91	.51	.07	.10	.52	.56	**.59**	.48	.48	.47	.48	.35	.54	.39	.34	.36
>= others	.96	.66	.16	.17	.67	**.723**	**.723**	.63	.60	.61	.60	.51	.66	.51	.47	.50

Table 23: Ranking scores for entries in the Spanish-English task (individual system track).

	REF	CEU-UPV	CU-ZEMAN	KOC	ONLINE-A	ONLINE-B	PROMT	RBMT-1	RBMT-2	RBMT-3	RBMT-4	RBMT-5	UEDIN	UOW	UPM	UPPSALA
REF	–	.06‡	.03‡	.09‡	.09‡	.09‡	.05‡	.03‡	.06‡	.04‡	.08‡	.02‡	.08‡	.02‡	.03‡	.04‡
CEU-UPV	**.84**‡	–	.21‡	.20†	**.43**	**.36**	**.42**	.37	.34⋆	**.50**†	.31	.34	**.32**	.21†	.13‡	.22
CU-ZEMAN	**.87**‡	**.56**‡	–	**.38**⋆	**.56**‡	**.56**‡	**.58**‡	**.46**⋆	**.40**	**.70**‡	**.46**⋆	**.49**†	**.51**‡	**.45**‡	.19‡	**.49**‡
KOC	**.84**‡	**.41**†	.22⋆	–	**.56**‡	**.51**‡	**.48**†	**.54**‡	**.39**	**.55**‡	**.42**	.35	**.51**‡	**.44**	.11‡	**.34**
ONLINE-A	**.72**‡	.31	.24‡	.15‡	–	**.36**	.37	.28†	.23‡	.35	.25‡	.20‡	.29⋆	.25†	.08‡	.09‡
ONLINE-B	**.72**‡	.30	.17‡	.18‡	.26	–	.29	.23‡	.20‡	**.37**	.20‡	.19‡	.19‡	.22‡	.02‡	.23⋆
PROMT	**.76**‡	.29	.21‡	.25†	**.42**	**.43**	–	.24‡	.24	.19	.27⋆	.26†	.32	.25‡	.18‡	.21‡
RBMT-1	**.85**‡	.37	.29⋆	.23‡	**.51**†	**.54**‡	**.48**‡	–	.35	**.45**‡	**.40**†	.05‡	**.47**	.39	.25‡	.39
RBMT-2	**.86**‡	**.50**⋆	.35	.38	**.51**‡	**.48**‡	**.35**	**.39**	–	**.41**†	.34	.36	**.45**	.36	.23‡	**.41**
RBMT-3	**.86**‡	.26†	.18‡	.22‡	**.40**	.35	.19	.20‡	.22†	–	.25†	.23‡	.24‡	.33	.10‡	.22†
RBMT-4	**.80**‡	**.45**	.29⋆	.34	**.53**‡	**.51**‡	**.43**⋆	.21†	**.38**	**.43**†	–	.24‡	.34	.30	.20‡	**.45**⋆
RBMT-5	**.96**‡	**.43**	.29†	**.42**	**.57**‡	**.61**‡	**.46**†	**.22**‡	**.38**	**.49**‡	**.47**‡	–	**.50**	**.46**	.27†	**.47**
UEDIN	**.74**‡	.28	.20‡	.21‡	**.46**⋆	**.48**‡	**.43**	.37	.31	**.49**‡	**.45**	.35	–	.20†	.14‡	.23
UOW	**.90**‡	**.44**†	.18‡	.32	**.46**†	**.52**‡	**.56**‡	.39	**.39**	**.44**	**.45**	.36	**.38**†	–	.10‡	**.32**
UPM	**.93**‡	**.65**‡	**.53**‡	**.67**‡	**.74**‡	**.71**‡	**.69**‡	**.59**‡	**.51**‡	**.74**‡	**.60**‡	**.51**†	**.64**‡	**.68**‡	–	**.62**‡
UPPSALA	**.84**‡	**.36**	.21‡	.32	**.49**‡	**.42**⋆	**.45**‡	.39	.35	**.45**†	.29⋆	.41	**.35**	.30	.15‡	–
> others	.83	.38	.24	.30	**.47**	.46	.41	.33	.32	.43	.35	.29	.38	.33	.14	.31
>= others	.94	.65	.49	.56	.72	**.74**	.70	.60	.57	.71	.61	.54	.64	.59	.34	.61

Table 24: Ranking scores for entries in the English-Spanish task (individual system track).

	REF	CMU-DENKOWSKI	CMU-HANNEMAN	CU-ZEMAN	JHU	KIT	LIA-LIG	LIMSI	LIUM	ONLINE-A	ONLINE-B	RBMT-1	RBMT-2	RBMT-3	RBMT-4	RBMT-5	RWTH-HUCK	SYSTRAN	UEDIN
REF	–	.10‡	.18‡	.06‡	.03‡	.14‡	.15‡	.14‡	.14‡	.12‡	.05‡	.12‡	.09‡	.05‡	.06‡	.05‡	.05‡	.07‡	.02‡
CMU-DENKOWSKI	**.79**‡	–	.35	.12‡	.34	.32	**.41**	**.35**	.21⋆	**.47**⋆	**.46**	**.49**	.32	.33	.36	.35	.25	**.45**	.29
CMU-HANNEMAN	**.79**‡	.35	–	.17‡	.29	**.44**⋆	**.43**	**.52**⋆	**.45**	**.45**	**.49**	**.51**	**.39**	**.44**	**.38**	.35	.35	**.43**	.37
CU-ZEMAN	**.94**‡	**.61**‡	**.67**‡	–	**.54**†	**.66**‡	**.66**†	**.58**†	**.60**‡	**.59**†	**.88**‡	**.62**‡	**.59**⋆	**.63**‡	**.60**†	**.56**	**.68**‡	**.64**†	.40
JHU	**.82**‡	.34	.29	.22†	–	.26	**.54**⋆	**.40**	.36	**.43**	.40	**.49**	**.42**	.40	.34	.35	.36	**.47**	.20†
KIT	**.79**‡	**.39**	.20⋆	.16‡	**.40**	–	.26⋆	**.46**	.34	**.38**	**.52**	.38	.35	.39	.28	.38	.15†	.32	.30
LIA-LIG	**.75**‡	.24	.31	.28†	.24⋆	**.59**⋆	–	**.49**	.27	**.40**	**.46**	.35	.26	.31⋆	.29	.32	.32	.33⋆	.35
LIMSI	**.86**‡	.30	.25⋆	.21†	.31	.26	.26	–	.38	**.40**	**.42**	.35	.18†	.43	.34	.16‡	.34	.34	.33
LIUM	**.78**‡	**.45**⋆	.33	.16‡	**.38**	.34	**.44**	**.40**	–	.38	.30	.44	.26†	.33⋆	.38	.28	.29	.33	.28
ONLINE-A	**.80**‡	.23⋆	.21	.22†	.37	.35	.36	.33	**.46**	–	**.43**	.35	.16‡	.33	.24†	.20‡	.26	.34	.27†
ONLINE-B	**.86**‡	.37	.31	.04‡	**.46**	.22	.36	.33	**.43**	.26	–	.40	.20†	.16‡	**.44**	.20‡	.41	.38	.22†
RBMT-1	**.87**‡	.44	.35	.23‡	.46	**.44**	**.54**	**.48**	.44	**.53**	**.54**	–	.39	.37	.33	.11†	.39	.17†	.35
RBMT-2	**.84**‡	**.47**	.37	.26⋆	.40	**.50**	**.45**	**.52**†	**.54**†	**.58**‡	**.67**†	**.45**	–	**.51**	.35	.22†	**.51**	**.57**	**.41**
RBMT-3	**.89**‡	**.44**	.42	.19‡	.40	**.43**	**.54**⋆	**.46**	**.61**⋆	**.50**	**.71**‡	.37	.32	–	.42	.35	.42	**.47**	.40
RBMT-4	**.85**‡	**.53**	.36	.26†	**.51**	**.47**	**.55**	**.52**	**.46**	**.59**†	.40	**.43**	**.50**	.42	–	.34	**.46**	.44	**.41**
RBMT-5	**.93**‡	**.58**	**.55**	.33	**.54**	**.54**	**.59**	**.70**‡	**.56**	**.66**‡	**.65**‡	**.36**†	**.54**†	**.46**	**.37**	–	**.50**	**.54**⋆	**.54**
RWTH-HUCK	**.92**‡	**.43**	**.38**	.14‡	.36	**.59**†	**.41**	**.44**	.29	**.53**	**.48**	**.46**	.30	**.46**	.32	.38	–	.37	.17†
SYSTRAN	**.93**‡	.39	.38	.24†	.44	**.48**	**.60**⋆	**.50**	**.40**	**.55**	**.57**	**.45**†	.36	.29	.44	.21⋆	**.49**	–	.36
UEDIN	**.93**‡	**.48**	**.41**	.40	**.51**†	**.48**	**.54**	**.49**	**.46**	**.60**†	**.57**†	**.52**	.37	**.47**	.39	.39	**.51**†	**.52**	–
> others	.85	.39	.36	.21	.39	.41	.46	.46	.41	.46	**.50**	.41	.33	.39	.35	.28	.37	.39	.32
>= others	.91	.62	.58	.37	.61	.64	.64	**.661**	.63	**.661**	.66	.58	.52	.55	.53	.45	.58	.54	.50

Table 25: Ranking scores for entries in the French-English task (individual system track).

	REF	CU-ZEMAN	JHU	KIT	LATL-GENEVA	LIMSI	LIUM	ONLINE-A	ONLINE-B	RBMT-1	RBMT-2	RBMT-3	RBMT-4	RBMT-5	RWTH-HUCK	UEDIN	UPPSALA	UPPSALA-FBK
REF	–	.07‡	.06‡	.25‡	.07‡	.13‡	.20‡	.15‡	.20‡	.10‡	.09‡	.18‡	.11‡	.12‡	.14‡	.18‡	.16‡	.16‡
CU-ZEMAN	**.92**‡	–	**.83**‡	**.86**‡	**.63**†	**.85**‡	**.90**‡	**.86**‡	**.81**‡	**.89**‡	**.70**‡	**.75**‡	**.75**‡	**.61**‡	**.78**‡	**.79**‡	**.81**‡	**.81**‡
JHU	**.91**‡	.07‡	–	**.55**†	.30⋆	**.60**‡	**.50**⋆	**.55**⋆	**.59**‡	**.45**	.41	.34⋆	.30†	**.50**	**.40**	**.42**	**.42**	**.44**
KIT	**.63**‡	.04‡	.29†	–	.18‡	**.47**	.37	.30⋆	.37	.38	.30†	.37	.24‡	.34	.28	.34	.24†	.13‡
LATL-GENEVA	**.86**‡	.29†	**.54**⋆	**.73**‡	–	**.77**‡	**.67**‡	**.71**‡	**.79**‡	**.55**†	**.39**	**.66**‡	**.52**	**.58**‡	**.58**‡	**.51**	**.52**	**.58**†
LIMSI	**.75**‡	.04‡	.21‡	.29	.13‡	–	.23⋆	.28⋆	.37	.27†	.27‡	.24‡	.24‡	.21‡	.27†	.28⋆	.25†	.31
LIUM	**.76**‡	.04‡	.26⋆	**.44**	.24‡	**.46**⋆	–	.33	**.52**	**.48**	.25‡	.36	.25‡	.28†	**.43**	**.40**	.35	.32
ONLINE-A	**.78**‡	.10‡	.31⋆	**.51**⋆	.22‡	**.51**⋆	**.46**	–	**.44**	.39	.36	.41	.30⋆	.41	**.41**	.32⋆	.46	.33
ONLINE-B	**.70**‡	.06‡	.27‡	**.41**	.13‡	**.39**	.32	.30	–	**.47**	.22‡	.26†	.13‡	.28†	.32	.26†	.33	.27†
RBMT-1	**.83**‡	.07‡	.38	**.46**	.23†	**.56**†	.39	**.41**	.42	–	.17‡	.34	.36	.13	**.52**	.33⋆	.40	.40
RBMT-2	**.88**‡	.25‡	**.47**	**.59**†	.37	**.65**‡	**.63**‡	**.51**	**.57**‡	**.54**‡	–	**.58**‡	.39	**.54**⋆	**.63**‡	**.61**†	**.47**	.42
RBMT-3	**.80**‡	.19‡	**.54**⋆	**.42**	.20‡	**.60**‡	**.47**	**.44**	**.52**†	**.42**	.18‡	–	.21†	**.43**	**.51**	**.55**	.41	.39
RBMT-4	**.82**‡	.22‡	**.54**†	**.63**‡	.33	**.63**‡	**.64**‡	**.54**⋆	**.59**‡	**.41**	**.44**	**.46**†	–	**.47**	**.68**‡	**.53**	**.42**	.39
RBMT-5	**.86**‡	.18‡	.46	**.53**	.20‡	**.62**‡	**.56**†	**.46**	**.61**†	**.22**	.33⋆	.40	.34	–	**.43**	**.52**	.40	**.53**⋆
RWTH-HUCK	**.76**‡	.08‡	.33	**.38**	.21‡	**.60**†	.40	.38	**.43**	.36	.18‡	.37	.21‡	.38	–	.39	.22‡	.29
UEDIN	**.78**‡	.15‡	.37	**.46**	.34	**.49**⋆	.38	**.53**⋆	**.58**†	**.56**⋆	.33†	.35	.36	.37	**.47**	–	.38	.31
UPPSALA	**.77**‡	.07‡	.36	**.53**†	.36	**.49**†	**.46**	.46	**.56**	**.46**	.38	**.42**	.39	**.55**	**.57**‡	**.39**	–	**.47**
UPPSALA-FBK	**.80**‡	.10‡	.40	**.71**‡	.27†	**.50**	**.47**	**.51**	**.53**†	**.42**	**.48**	**.41**	**.52**	.29⋆	**.50**	**.47**	.40	–
> others	.80	.12	.39	.51	.25	**.55**	.48	.45	.52	.43	.32	.41	.33	.39	.46	.43	.39	.38
>= others	.86	.20	.55	.69	.39	**.73**	.64	.60	.70	.61	.46	.58	.49	.55	.65	.58	.55	.54

Table 26: Ranking scores for entries in the English-French task (individual system track).

	REF	BM-I2R	CMU-DENKOWSKI	CMU-HEWAVITHARANA	HYDERABAD	KOC	LIU	UMD-EIDELMAN	UMD-HU	UPPSALA
REF	–	.03‡	.01‡	.03‡	.02‡	.01‡	.00‡	.01‡	.01‡	.02‡
BM-I2R	**.91**‡	–	.28†	.27†	.13‡	.08‡	.19‡	.30†	.30‡	.24‡
CMU-DENKOWSKI	**.93**‡	**.44**†	–	.25	.22‡	.15‡	.28†	.33	.29‡	.31†
CMU-HEWAVITHARANA	**.91**‡	**.40**†	**.31**	–	.21‡	.16‡	.29†	.35	.39	.30
HYDERABAD	**.96**‡	**.71**‡	**.59**‡	**.58**‡	–	.27‡	**.56**‡	**.57**‡	**.42**	**.52**‡
KOC	**.94**‡	**.78**‡	**.75**‡	**.64**‡	**.55**‡	–	**.65**‡	**.69**‡	**.62**‡	**.64**‡
LIU	**.92**‡	**.56**‡	**.42**†	**.44**†	.27‡	.24‡	–	**.43**	.41	.39
UMD-EIDELMAN	**.94**‡	**.44**†	**.35**	.35	.17‡	.17‡	.34	–	.37	.31⋆
UMD-HU	**.90**‡	**.50**‡	**.57**‡	**.45**	.35	.21‡	**.46**	**.45**	–	**.42**
UPPSALA	**.93**‡	**.48**‡	**.47**†	**.39**	.31‡	.20‡	**.40**	**.43**⋆	.37	–
> others	.93	**.49**	.42	.39	.25	.17	.35	.40	.36	.35
>= others	.98	**.71**	.66	.64	.43	.31	.55	.63	.52	.57

Table 27: Ranking scores for entries in the Haitian Creole (Clean)-English task (individual system track).

	REF	BM-I2R	CMU-DENKOWSKI	CMU-HEWAVITHARANA	JHU	LIU	UMD-EIDELMAN
REF	–	.05‡	.03‡	.04‡	.02‡	.02‡	.03‡
BM-I2R	**.83**‡	–	.29†	.25‡	.22‡	.30‡	.30‡
CMU-DENKOWSKI	**.89**‡	**.44**†	–	**.37**⋆	.23‡	.37	.30†
CMU-HEWAVITHARANA	**.86**‡	**.43**‡	.26⋆	–	.27‡	**.37**	.32
JHU	**.96**‡	**.62**‡	**.53**‡	**.49**‡	–	**.52**‡	**.47**‡
LIU	**.92**‡	**.48**‡	**.38**	.34	.31‡	–	.36
UMD-EIDELMAN	**.92**‡	**.48**‡	**.44**†	**.42**	.29‡	**.41**	–
> others	.90	**.43**	.34	.33	.23	.34	.30
>= others	.97	**.65**	.59	.60	.41	.55	.52

Table 28: Ranking scores for entries in the Haitian Creole (Raw)-English task (individual system track).

	REF	BBN-COMBO	CMU-HEAFIELD-COMBO	JHU-COMBO	UPV-PRHLT-COMBO
REF	–	.01‡	.02‡	.01‡	.01‡
BBN-COMBO	**.91**‡	–	**.25**	.18⋆	.16‡
CMU-HEAFIELD-COMBO	**.90**‡	.24	–	.17‡	.12‡
JHU-COMBO	**.92**‡	**.27**⋆	**.29**‡	–	.20‡
UPV-PRHLT-COMBO	**.94**‡	**.41**‡	**.42**‡	**.36**‡	–
> others	.92	.23	**.24**	.18	.12
>= others	.99	.62	**.64**	.58	.47

Table 29: Ranking scores for entries in the Czech-English task (system combination track).

	REF	CMU-HEAFIELD-COMBO	UPV-PRHLT-COMBO
REF	–	.04‡	.04‡
CMU-HEAFIELD-COMBO	**.86**‡	–	.17‡
UPV-PRHLT-COMBO	**.88**‡	**.30**‡	–
> others	.87	**.17**	.11
>= others	.96	**.48**	.41

Table 30: Ranking scores for entries in the English-Czech task (system combination track).

	REF	BBN-COMBO	CMU-HEAFIELD-COMBO	JHU-COMBO	KOC-COMBO	QUAERO-COMBO	RWTH-LEUSCH-COMBO	UPV-PRHLT-COMBO	UZH-COMBO
REF	–	.11‡	.09‡	.04‡	.09‡	.10‡	.14‡	.05‡	.09‡
BBN-COMBO	**.79**‡	–	**.45**‡	.32	.21‡	.28†	**.39**	.31⋆	**.36**
CMU-HEAFIELD-COMBO	**.84**‡	.23‡	–	.21‡	.17‡	.19‡	.25⋆	.19‡	.31
JHU-COMBO	**.85**‡	**.42**	**.55**‡	–	.25†	.28†	**.40**†	.28⋆	**.47**⋆
KOC-COMBO	**.83**‡	**.56**‡	**.62**‡	**.45**†	–	**.41**	**.54**‡	**.40**⋆	**.51**†
QUAERO-COMBO	**.86**‡	**.52**†	**.64**‡	**.45**†	.36	–	**.54**‡	**.49**†	**.48**
RWTH-LEUSCH-COMBO	**.83**‡	.28	**.41**⋆	.22†	.20‡	.22‡	–	.22‡	.38
UPV-PRHLT-COMBO	**.85**‡	**.47**⋆	**.57**‡	**.42**⋆	.25⋆	.26†	**.48**‡	–	**.49**†
UZH-COMBO	**.86**‡	.34	**.38**	.31⋆	.29†	.32	**.41**	.30†	–
> others	.84	.36	**.46**	.30	.22	.26	.39	.27	.39
>= others	.91	.61	**.70**	.56	.45	.46	.65	.52	.60

Table 31: Ranking scores for entries in the German-English task (system combination track).

	REF	CMU-HEAFIELD-COMBO	KOC-COMBO	UPV-PRHLT-COMBO	UZH-COMBO
REF	–	.11‡	.09‡	.10‡	.11‡
CMU-HEAFIELD-COMBO	**.81**‡	–	.19‡	.23‡	.32
KOC-COMBO	**.84**‡	**.48**‡	–	**.38**‡	**.47**‡
UPV-PRHLT-COMBO	**.81**‡	**.36**‡	.23‡	–	**.37**⋆
UZH-COMBO	**.80**‡	**.34**	.24‡	.31⋆	–
> others	.81	**.320**	.19	.25	**.318**
>= others	.90	**.61**	.46	.56	.58

Table 32: Ranking scores for entries in the English-German task (system combination track).

	REF	BBN-COMBO	CMU-HEAFIELD-COMBO	JHU-COMBO	KOC-COMBO	RWTH-LEUSCH-COMBO	UPV-PRHLT-COMBO
REF	–	.05‡	.09‡	.05‡	.07‡	.06‡	.08‡
BBN-COMBO	**.81**‡	–	**.34**	**.27**	.21‡	**.27**	.26
CMU-HEAFIELD-COMBO	**.84**‡	.31	–	.18‡	.15‡	**.29**	.20
JHU-COMBO	**.83**‡	.25	**.32**‡	–	.27	**.35**‡	.25
KOC-COMBO	**.84**‡	**.39**‡	**.39**‡	**.32**	–	**.39**‡	**.31**⋆
RWTH-LEUSCH-COMBO	**.81**‡	.24	.23	.16‡	.17‡	–	.14‡
UPV-PRHLT-COMBO	**.77**‡	**.30**	**.26**	**.27**	.22⋆	**.35**‡	–
> others	.82	.25	.27	.21	.18	**.28**	.21
>= others	.93	.64	.67	.62	.56	**.71**	.64

Table 33: Ranking scores for entries in the Spanish-English task (system combination track).

	REF	CMU-HEAFIELD-COMBO	KOC-COMBO	UOW-COMBO	UPV-PRHLT-COMBO
REF	–	.10‡	.07‡	.09‡	.08‡
CMU-HEAFIELD-COMBO	**.70**‡	–	.15‡	.21‡	.17‡
KOC-COMBO	**.76**‡	**.35**‡	–	**.36**‡	**.19**
UOW-COMBO	**.72**‡	**.29**‡	.22‡	–	.25‡
UPV-PRHLT-COMBO	**.76**‡	**.35**‡	.16	**.35**‡	–
> others	.73	**.27**	.15	.25	.17
>= others	.91	**.69**	.58	.63	.59

Table 34: Ranking scores for entries in the English-Spanish task (system combination track).

	REF	BBN-COMBO	CMU-HEAFIELD-COMBO	JHU-COMBO	LIUM-COMBO	RWTH-LEUSCH-COMBO	UPV-PRHLT-COMBO
REF	–	.04‡	.04‡	.06‡	.06‡	.06‡	.02‡
BBN-COMBO	**.82**‡	–	**.35**	.25	.18‡	.21⋆	.21‡
CMU-HEAFIELD-COMBO	**.90**‡	.29	–	.30	.20‡	.29	.25†
JHU-COMBO	**.83**‡	**.35**	**.40**	–	.31⋆	**.36**	.21†
LIUM-COMBO	**.83**‡	**.42**‡	**.40**‡	**.44**⋆	–	**.38**†	**.35**
RWTH-LEUSCH-COMBO	**.83**‡	**.34**⋆	.29	.30	.22†	–	.21‡
UPV-PRHLT-COMBO	**.91**‡	**.49**‡	**.40**†	**.34**†	.30	**.40**‡	–
> others	.85	**.32**	.31	.28	.21	.28	.21
>= others	.95	**.67**	.62	.59	.53	.63	.53

Table 35: Ranking scores for entries in the French-English task (system combination track).

	REF	CMU-HEAFIELD-COMBO	UPV-PRHLT-COMBO
REF	–	.11‡	.11‡
CMU-HEAFIELD-COMBO	**.74**‡	–	.23‡
UPV-PRHLT-COMBO	**.77**‡	**.38**‡	–
> others	.76	**.24**	.17
>= others	.89	**.51**	.43

Table 36: Ranking scores for entries in the English-French task (system combination track).

	REF	CMU-HEAFIELD-COMBO	KOC-COMBO	UPV-PRHLT-COMBO
REF	–	.01‡	.01‡	.01‡
CMU-HEAFIELD-COMBO	**.94**‡	–	.29‡	.21‡
KOC-COMBO	**.96**‡	**.48**‡	–	**.41**†
UPV-PRHLT-COMBO	**.94**‡	**.34**‡	.29†	–
> others	.95	**.28**	.20	.21
>= others	.99	**.52**	.38	.48

Table 37: Ranking scores for entries in the Haitian Creole (Clean)-English task (system combination track).

	REF	CMU-HEAFIELD-COMBO	UPV-PRHLT-COMBO
REF	–	.02‡	.02‡
CMU-HEAFIELD-COMBO	**.83**‡	–	.24
UPV-PRHLT-COMBO	**.86**‡	**.29**	–
> others	.84	**.16**	.13
>= others	.98	**.47**	.43

Table 38: Ranking scores for entries in the Haitian Creole (Raw)-English task (system combination track).

	AMBER	AMBER-NL	AMBER-TI	BLEU	F15	F15G3	MTERATER	MTERATER-PLUS	ROSE	TER	TINE-SRL-MATCH	METEOR-1.3-ADQ	METEOR-1.3-RANK	MP4IBM1	MPF	TESLA-B	TESLA-F	TESLA-M	WMPF
Czech-English News Task																			
BBN-COMBO	0.24	0.24	0.25	0.29	0.31	0.19	–9627	–10667	1.97	0.53	0.49	0.61	0.34	–65	44	0.48	0.03	0.51	43
CMU-HEAFIELD-COMBO	0.24	0.24	0.24	0.28	0.3	0.18	–9604	–10933	1.97	0.54	0.5	0.60	0.33	–65	43	0.48	0.03	0.52	42
CST	0.19	0.19	0.2	0.16	0.21	0.10	–27410	–27880	1.94	0.64	0.40	0.5	0.28	–65	34	0.38	0.02	0.42	33
CU-BOJAR	0.21	0.21	0.22	0.19	0.24	0.13	–23441	–22289	1.95	0.64	0.44	0.55	0.30	–65	37	0.42	0.02	0.46	36
CU-ZEMAN	0.20	0.2	0.21	0.14	0.21	0.11	–33520	–30938	1.93	0.66	0.38	0.52	0.29	–66	31	0.37	0.02	0.40	30
JHU	0.22	0.21	0.22	0.2	0.25	0.13	–21278	–20480	1.95	0.60	0.43	0.55	0.30	–65	37	0.42	0.02	0.46	36
JHU-COMBO	0.24	0.23	0.24	0.29	0.31	0.19	–12563	–12688	1.97	0.53	0.5	0.60	0.33	–65	44	0.48	0.03	0.52	43
ONLINE-B	0.24	0.23	0.24	0.29	0.31	0.19	–10673	–11506	1.97	0.52	0.50	0.60	0.33	–65	44	0.49	0.03	0.52	43
SYSTRAN	0.20	0.2	0.21	0.18	0.22	0.11	–23996	–24570	1.94	0.63	0.42	0.52	0.29	–65	36	0.4	0.02	0.45	34
UEDIN	0.22	0.22	0.23	0.22	0.26	0.14	–14958	–15342	1.96	0.59	0.45	0.57	0.31	–65	40	0.44	0.03	0.48	39
UPPSALA	0.21	0.20	0.21	0.20	0.23	0.12	–22233	–22509	1.95	0.62	0.43	0.53	0.29	–65	37	0.41	0.02	0.46	36
UPV-PRHLT-COMBO	0.24	0.23	0.24	0.29	0.31	0.19	–13904	–15260	1.97	0.54	0.49	0.60	0.33	–65	44	0.48	0.03	0.52	43

Table 39: Automatic evaluation metric scores for systems in the WMT11 Czech-English News Task (newssyscombtest2011)

	AMBER	AMBER-NL	AMBER-TI	BLEU	F15	F15G3	MTERATER	MTERATER-PLUS	ROSE	TER	TINE-SRL-MATCH	DFKI-PARSECONF	METEOR-1.3-ADQ	METEOR-1.3-RANK	MP4IBM1	MPF	TESLA-B	TESLA-F	TESLA-M	WMPF
German-English News Task																				
BBN-COMBO	0.23	0.22	0.23	0.25	0.28	0.16	–17103	–17837	1.97	0.56	0.46	0.06	0.59	0.32	–43	42	0.46	0.03	0.49	41
CMU-DYER	0.21	0.21	0.22	0.22	0.25	0.13	–26089	–29214	1.95	0.59	0.44	0.04	0.56	0.31	–45	39	0.43	0.03	0.47	38
CMU-HEAFIELD-COMBO	0.23	0.22	0.23	0.24	0.27	0.15	–12868	–16156	1.96	0.57	0.47	0.07	0.58	0.32	–44	41	0.46	0.03	0.51	40
CST	0.19	0.18	0.19	0.17	0.22	0.11	–61131	–60157	1.94	0.63	0.39	0.03	0.5	0.27	–46	34	0.37	0.02	0.41	33
CU-ZEMAN	0.2	0.19	0.20	0.14	0.22	0.11	–64860	–61329	1.93	0.65	0.37	0.06	0.51	0.28	–47	31	0.37	0.02	0.4	30
DFKI-XU	0.21	0.20	0.21	0.21	0.25	0.14	–40171	–39455	1.95	0.58	0.44	0.03	0.54	0.3	–45	38	0.42	0.02	0.46	37
JHU	0.19	0.19	0.2	0.17	0.22	0.11	–62997	–58673	1.94	0.64	0.39	0.03	0.51	0.28	–45	34	0.38	0.02	0.41	33
JHU-COMBO	0.22	0.22	0.23	0.24	0.27	0.15	–30492	–27016	1.96	0.57	0.46	0.04	0.57	0.31	–44	41	0.45	0.03	0.48	39
KIT	0.21	0.21	0.22	0.22	0.25	0.13	–31064	–31930	1.95	0.6	0.44	0.05	0.55	0.31	–44	39	0.43	0.02	0.47	37
KOC	0.2	0.2	0.20	0.18	0.23	0.12	–52337	–50231	1.94	0.63	0.41	0.05	0.52	0.29	–45	35	0.39	0.02	0.43	34
KOC-COMBO	0.21	0.21	0.21	0.22	0.26	0.14	–40002	–38374	1.96	0.59	0.44	0.03	0.54	0.3	–44	38	0.42	0.02	0.46	37
LIMSI	0.21	0.20	0.21	0.20	0.24	0.13	–39419	–38297	1.95	0.61	0.43	0.04	0.54	0.3	–44	38	0.42	0.02	0.46	36
LINGUATEC	0.19	0.19	0.2	0.16	0.22	0.11	–26064	–31116	1.94	0.68	0.42	0.15	0.53	0.29	–46	35	0.42	0.02	0.47	34
LIU	0.21	0.20	0.21	0.2	0.24	0.13	–40281	–40496	1.95	0.62	0.43	0.04	0.53	0.29	–44	37	0.41	0.02	0.45	36
ONLINE-A	0.22	0.21	0.22	0.21	0.26	0.14	–25411	–25675	1.95	0.6	0.45	0.06	0.57	0.31	–44	39	0.45	0.03	0.48	38
ONLINE-B	0.22	0.22	0.23	0.23	0.27	0.15	–15149	–19578	1.96	0.58	0.46	0.06	0.57	0.32	–44	41	0.46	0.03	0.5	39
QUAERO-COMBO	0.21	0.21	0.22	0.22	0.26	0.14	–34486	–33449	1.96	0.58	0.45	0.03	0.55	0.30	–44	39	0.43	0.03	0.47	38
RBMT-1	0.20	0.2	0.21	0.16	0.21	0.11	–32960	–34972	1.94	0.67	0.42	0.08	0.52	0.29	–45	36	0.42	0.02	0.46	34
RBMT-2	0.19	0.19	0.2	0.15	0.2	0.1	–40842	–43413	1.94	0.69	0.4	0.11	0.50	0.28	–45	34	0.4	0.02	0.44	33
RBMT-3	0.20	0.2	0.21	0.17	0.22	0.11	–32476	–33417	1.94	0.65	0.42	0.09	0.53	0.29	–44	36	0.42	0.02	0.47	35
RBMT-4	0.20	0.2	0.21	0.17	0.22	0.11	–34287	–34604	1.94	0.66	0.42	0.08	0.52	0.29	–45	36	0.42	0.02	0.47	35
RBMT-5	0.19	0.19	0.20	0.15	0.20	0.10	–49097	–46635	1.94	0.68	0.40	0.07	0.50	0.28	–46	34	0.4	0.02	0.44	33
RWTH-LEUSCH-COMBO	0.22	0.22	0.23	0.24	0.28	0.16	–22878	–22089	1.96	0.56	0.46	0.03	0.58	0.32	–44	41	0.45	0.03	0.49	40
RWTH-WUEBKER	0.21	0.20	0.21	0.21	0.24	0.13	–35973	–37140	1.95	0.60	0.44	0.04	0.54	0.3	–45	38	0.42	0.02	0.45	37
UEDIN	0.21	0.20	0.21	0.19	0.23	0.12	–32791	–34633	1.95	0.63	0.43	0.07	0.54	0.3	–45	37	0.42	0.02	0.46	36
UPPSALA	0.20	0.2	0.21	0.2	0.23	0.12	–40448	–41548	1.95	0.63	0.42	0.06	0.53	0.29	–45	37	0.41	0.02	0.44	36
UPV-PRHLT-COMBO	0.22	0.21	0.22	0.23	0.27	0.15	–33413	–31778	1.96	0.58	0.45	0.03	0.57	0.31	–44	40	0.44	0.03	0.48	39
UZH-COMBO	0.22	0.21	0.22	0.23	0.27	0.15	–16326	–20831	1.96	0.58	0.45	0.07	0.57	0.31	–44	40	0.45	0.03	0.48	39

Table 40: Automatic evaluation metric scores for systems in the WMT11 German-English News Task (newssyscombtest2011)

	AMBER	AMBER-NL	AMBER-TI	BLEU	F15	F15G3	MTERATER	MTERATER-PLUS	ROSE	TER	TINE-SRL-MATCH	METEOR-1.3-ADQ	METEOR-1.3-RANK	MP4IBM1	MPF	TESLA-B	TESLA-F	TESLA-M	WMPF
French-English News Task																			
BBN-COMBO	0.25	0.25	0.26	0.31	0.32	0.21	–19552	–22107	1.98	0.48	0.51	0.64	0.36	–43	47	0.49	0.03	0.54	46
CMU-DENKOWSKI	0.24	0.24	0.24	0.26	0.29	0.17	–34357	–37807	1.97	0.53	0.48	0.61	0.34	–45	43	0.46	0.03	0.50	42
CMU-HANNEMAN	0.24	0.23	0.24	0.27	0.29	0.17	–33662	–37698	1.97	0.52	0.49	0.60	0.33	–45	44	0.46	0.03	0.51	42
CMU-HEAFIELD-COMBO	0.25	0.25	0.25	0.30	0.31	0.2	–18365	–22937	1.98	0.5	0.51	0.63	0.35	–44	46	0.49	0.03	0.54	45
CU-ZEMAN	0.22	0.22	0.23	0.17	0.24	0.13	–67586	–64688	1.94	0.6	0.41	0.56	0.31	–47	34	0.39	0.02	0.42	33
JHU	0.24	0.24	0.24	0.25	0.29	0.17	–41567	–39578	1.96	0.53	0.47	0.61	0.34	–45	42	0.46	0.03	0.5	41
JHU-COMBO	0.25	0.25	0.25	0.31	0.32	0.20	–32785	–31712	1.98	0.49	0.50	0.63	0.35	–43	47	0.48	0.03	0.53	45
KIT	0.25	0.24	0.25	0.29	0.31	0.19	–22678	–28283	1.98	0.51	0.50	0.63	0.35	–44	46	0.49	0.03	0.53	44
LIA-LIG	0.25	0.24	0.25	0.29	0.3	0.18	–34063	–34716	1.97	0.52	0.49	0.62	0.34	–44	45	0.48	0.03	0.52	44
LIMSI	0.25	0.24	0.25	0.28	0.29	0.18	–26269	–29363	1.97	0.52	0.5	0.62	0.34	–44	45	0.48	0.03	0.52	44
LIUM	0.25	0.24	0.25	0.29	0.30	0.19	–29288	–36137	1.98	0.52	0.49	0.62	0.34	–44	45	0.48	0.03	0.53	44
LIUM-COMBO	0.25	0.24	0.25	0.31	0.31	0.2	–30678	–35365	1.98	0.50	0.5	0.62	0.34	–44	46	0.48	0.03	0.53	45
ONLINE-A	0.25	0.24	0.25	0.27	0.3	0.18	–38761	–34096	1.97	0.52	0.49	0.62	0.34	–44	44	0.48	0.03	0.52	43
ONLINE-B	0.25	0.24	0.25	0.29	0.31	0.19	–19157	–25284	1.98	0.50	0.51	0.62	0.35	–45	46	0.49	0.03	0.54	44
RBMT-1	0.24	0.23	0.24	0.23	0.26	0.15	–49115	–39153	1.96	0.59	0.46	0.60	0.33	–43	42	0.46	0.03	0.51	41
RBMT-2	0.23	0.22	0.23	0.21	0.24	0.13	–59549	–50466	1.95	0.63	0.44	0.57	0.32	–43	40	0.43	0.02	0.48	39
RBMT-3	0.23	0.23	0.23	0.22	0.25	0.14	–52047	–45073	1.96	0.59	0.46	0.58	0.32	–44	41	0.45	0.02	0.50	40
RBMT-4	0.23	0.22	0.24	0.22	0.25	0.14	–54507	–42933	1.96	0.63	0.45	0.59	0.33	–43	40	0.44	0.02	0.49	39
RBMT-5	0.23	0.22	0.23	0.21	0.24	0.13	–55545	–48332	1.95	0.62	0.45	0.57	0.32	–44	40	0.44	0.02	0.49	38
RWTH-HUCK	0.24	0.24	0.25	0.28	0.3	0.18	–44018	–42549	1.97	0.52	0.49	0.61	0.34	–44	44	0.47	0.03	0.51	43
RWTH-LEUSCH-COMBO	0.26	0.25	0.26	0.31	0.32	0.20	–21914	–21746	1.98	0.49	0.51	0.64	0.35	–43	47	0.50	0.03	0.54	46
SYSTRAN	0.24	0.23	0.24	0.25	0.27	0.16	–34321	–40119	1.96	0.54	0.48	0.59	0.33	–44	43	0.46	0.03	0.51	41
UEDIN	0.23	0.23	0.24	0.25	0.27	0.16	–47202	–47955	1.96	0.56	0.47	0.59	0.33	–45	42	0.45	0.03	0.49	40
UPV-PRHLT-COMBO	0.25	0.25	0.26	0.31	0.32	0.20	–26947	–28689	1.98	0.5	0.51	0.63	0.35	–43	47	0.49	0.03	0.54	46

Table 41: Automatic evaluation metric scores for systems in the WMT11 French-English News Task (newssyscombtest2011)

	AMBER	AMBER-NL	AMBER-TI	BLEU	F15	F15G3	MTERATER	MTERATER-PLUS	ROSE	TER	TINE-SRL-MATCH	METEOR-1.3-ADQ	METEOR-1.3-RANK	MP4IBM1	MPF	TESLA-B	TESLA-F	TESLA-M	WMPF
Spanish-English News Task																			
ALACANT	0.24	0.23	0.24	0.27	0.28	0.17	–30135	–29622	1.97	0.53	0.46	0.61	0.34	–45	43	0.46	0.03	0.50	42
BBN-COMBO	0.25	0.25	0.25	0.32	0.33	0.21	–15284	–16192	1.98	0.48	0.5	0.64	0.35	–44	47	0.49	0.03	0.53	46
CMU-HEAFIELD-COMBO	0.25	0.25	0.25	0.32	0.31	0.20	–13456	–16113	1.98	0.5	0.5	0.64	0.35	–44	47	0.5	0.03	0.54	46
CU-ZEMAN	0.20	0.20	0.21	0.16	0.22	0.12	–49428	–48440	1.93	0.61	0.36	0.51	0.28	–49	32	0.35	0.02	0.38	31
HYDERABAD	0.20	0.20	0.21	0.17	0.21	0.11	–47754	–47059	1.94	0.61	0.39	0.50	0.28	–47	34	0.36	0.02	0.41	33
JHU-COMBO	0.25	0.25	0.25	0.32	0.32	0.20	–23939	–22685	1.98	0.49	0.49	0.63	0.35	–44	47	0.48	0.03	0.52	46
KOC	0.24	0.24	0.24	0.26	0.29	0.17	–22724	–25857	1.96	0.53	0.46	0.61	0.34	–45	42	0.46	0.03	0.49	41
KOC-COMBO	0.25	0.24	0.25	0.28	0.30	0.19	–22678	–22267	1.97	0.52	0.48	0.62	0.34	–44	44	0.48	0.03	0.52	43
ONLINE-A	0.25	0.24	0.25	0.28	0.3	0.18	–19017	–20120	1.97	0.52	0.48	0.63	0.35	–44	45	0.48	0.03	0.52	43
ONLINE-B	0.24	0.24	0.24	0.29	0.30	0.19	–11980	–18589	1.97	0.50	0.49	0.62	0.34	–45	45	0.49	0.03	0.53	44
RBMT-1	0.24	0.24	0.25	0.28	0.28	0.17	–31202	–26151	1.97	0.57	0.46	0.61	0.34	–44	45	0.47	0.03	0.51	43
RBMT-2	0.23	0.23	0.24	0.24	0.25	0.15	–35157	–31405	1.96	0.6	0.44	0.59	0.33	–44	42	0.44	0.02	0.49	41
RBMT-3	0.23	0.23	0.24	0.25	0.26	0.15	–28289	–26082	1.97	0.59	0.45	0.6	0.33	–43	43	0.46	0.03	0.51	42
RBMT-4	0.24	0.23	0.24	0.25	0.26	0.16	–27892	–25546	1.97	0.59	0.46	0.60	0.33	–43	43	0.46	0.03	0.52	42
RBMT-5	0.24	0.23	0.24	0.27	0.26	0.16	–36770	–31613	1.96	0.58	0.45	0.6	0.33	–45	43	0.45	0.03	0.50	42
RWTH-LEUSCH-COMBO	0.25	0.25	0.26	0.32	0.32	0.21	–15172	–15261	1.98	0.49	0.5	0.64	0.35	–43	48	0.50	0.03	0.54	47
SYSTRAN	0.24	0.23	0.24	0.27	0.28	0.17	–20129	–26051	1.97	0.53	0.47	0.60	0.33	–46	44	0.46	0.03	0.51	42
UEDIN	0.22	0.22	0.23	0.22	0.25	0.14	–25462	–31678	1.96	0.58	0.45	0.57	0.32	–47	40	0.44	0.03	0.48	39
UFAL-UM	0.23	0.22	0.23	0.23	0.24	0.14	–42123	–37765	1.96	0.60	0.43	0.58	0.32	–43	41	0.43	0.02	0.48	40
UPM	0.22	0.22	0.23	0.22	0.24	0.14	–39748	–38433	1.95	0.58	0.43	0.57	0.32	–45	40	0.42	0.02	0.46	38
UPV-PRHLT-COMBO	0.25	0.25	0.26	0.32	0.32	0.20	–16094	–17723	1.98	0.50	0.49	0.64	0.35	–43	47	0.5	0.03	0.54	46

Table 42: Automatic evaluation metric scores for systems in the WMT11 Spanish-English News Task (newssyscombtest2011)

	AMBER	AMBER-NL	AMBER-TI	BLEU	F15	F15G3	ROSE	TER	METEOR-1.3-RANK	MP4IBM1	MPF	WMPF
English-Czech News Task												
CMU-HEAFIELD-COMBO	0.2	0.19	0.20	0.19	0.22	0.12	2.03	0.62	0.24	–62	29	27
COMMERCIAL1	0.16	0.15	0.16	0.11	0.16	0.08	2.01	0.70	0.19	–65	22	21
COMMERCIAL2	0.12	0.10	0.13	0.09	0.15	0.06	2.00	0.73	0.18	–65	21	19
CU-BOJAR	0.18	0.17	0.18	0.16	0.2	0.1	2.02	0.65	0.23	–63	26	24
CU-MARECEK	0.18	0.17	0.18	0.16	0.2	0.1	2.02	0.65	0.22	–63	26	24
CU-POPEL	0.17	0.16	0.18	0.14	0.19	0.1	2.02	0.66	0.21	–64	25	23
CU-TAMCHYNA	0.18	0.17	0.18	0.15	0.2	0.1	2.02	0.65	0.22	–63	26	24
CU-ZEMAN	0.17	0.16	0.17	0.13	0.18	0.09	2.02	0.66	0.21	–63	23	22
JHU	0.18	0.18	0.18	0.16	0.21	0.11	2.02	0.63	0.22	–63	26	24
ONLINE-B	0.2	0.19	0.20	0.2	0.22	0.12	2.03	0.62	0.24	–63	29	27
UEDIN	0.19	0.18	0.19	0.17	0.21	0.11	2.03	0.63	0.23	–63	27	26
UPV-PRHLT-COMBO	0.2	0.19	0.20	0.20	0.23	0.13	2.03	0.61	0.24	–63	29	28

Table 43: Automatic evaluation metric scores for systems in the WMT11 English-Czech News Task (newssyscombtest2011)

	AMBER	AMBER-NL	AMBER-TI	BLEU	F15	F15G3	ROSE	TER	METEOR-1.3-RANK	MP4IBM1	MPF	TESLA-B	TESLA-F	TESLA-M	WMPF
English-German News Task															
CMU-HEAFIELD-COMBO	0.19	0.18	0.19	0.17	0.21	0.11	1.96	0.66	0.39	–46	36	0.41	0.03	0.45	35
COPENHAGEN	0.17	0.17	0.18	0.14	0.18	0.09	1.95	0.69	0.36	–47	33	0.38	0.02	0.42	32
CU-TAMCHYNA	0.17	0.17	0.18	0.11	0.18	0.09	1.94	0.70	0.36	–48	31	0.36	0.02	0.4	30
CU-ZEMAN	0.16	0.15	0.16	0.05	0.17	0.08	1.92	0.71	0.34	–51	25	0.31	0.02	0.34	25
DFKI-FEDERMANN	0.17	0.16	0.17	0.13	0.17	0.08	1.95	0.71	0.34	–47	33	0.38	0.03	0.44	32
DFKI-XU	0.18	0.17	0.18	0.15	0.19	0.1	1.96	0.68	0.37	–47	35	0.39	0.03	0.43	34
ILLC-UVA	0.15	0.14	0.15	0.12	0.18	0.08	1.95	0.68	0.33	–49	32	0.36	0.02	0.4	31
JHU	0.17	0.17	0.18	0.14	0.18	0.09	1.95	0.68	0.35	–47	33	0.37	0.02	0.42	32
KIT	0.18	0.17	0.18	0.15	0.19	0.09	1.96	0.68	0.37	–47	35	0.39	0.03	0.43	34
KOC	0.17	0.16	0.17	0.12	0.17	0.08	1.95	0.69	0.35	–47	32	0.36	0.02	0.40	31
KOC-COMBO	0.18	0.17	0.18	0.15	0.2	0.1	1.95	0.67	0.37	–47	34	0.38	0.02	0.42	33
LIMSI	0.18	0.17	0.18	0.15	0.19	0.09	1.96	0.67	0.36	–47	35	0.39	0.03	0.44	33
LIU	0.17	0.17	0.18	0.15	0.19	0.09	1.95	0.68	0.36	–47	34	0.38	0.02	0.43	33
ONLINE-A	0.18	0.17	0.18	0.15	0.19	0.09	1.96	0.67	0.37	–47	35	0.40	0.03	0.45	33
ONLINE-B	0.19	0.18	0.19	0.17	0.21	0.11	1.96	0.65	0.38	–46	36	0.42	0.03	0.46	35
RBMT-1	0.17	0.17	0.18	0.13	0.18	0.08	1.95	0.7	0.35	–46	34	0.39	0.03	0.45	33
RBMT-2	0.16	0.16	0.17	0.12	0.16	0.08	1.94	0.73	0.33	–47	32	0.37	0.03	0.43	31
RBMT-3	0.18	0.17	0.18	0.14	0.18	0.09	1.95	0.69	0.36	–46	35	0.39	0.03	0.46	34
RBMT-4	0.17	0.16	0.17	0.13	0.17	0.08	1.95	0.70	0.34	–47	33	0.38	0.03	0.45	32
RBMT-5	0.17	0.16	0.17	0.12	0.17	0.08	1.95	0.71	0.34	–47	33	0.38	0.03	0.44	32
RWTH-FREITAG	0.17	0.17	0.17	0.15	0.19	0.09	1.95	0.68	0.36	–47	34	0.37	0.02	0.41	33
UEDIN	0.17	0.17	0.18	0.14	0.18	0.09	1.95	0.69	0.36	–47	34	0.38	0.02	0.42	33
UOW	0.17	0.16	0.17	0.13	0.17	0.08	1.95	0.7	0.35	–47	33	0.37	0.02	0.42	32
UPPSALA	0.17	0.16	0.17	0.14	0.18	0.09	1.95	0.68	0.35	–47	33	0.37	0.02	0.42	32
UPV-PRHLT-COMBO	0.18	0.18	0.19	0.17	0.20	0.10	1.96	0.66	0.38	–46	36	0.4	0.03	0.44	35
UZH-COMBO	0.19	0.18	0.19	0.17	0.21	0.11	1.96	0.66	0.38	–46	36	0.40	0.03	0.44	35

Table 44: Automatic evaluation metric scores for systems in the WMT11 English-German News Task (newssyscombtest2011)

	AMBER	AMBER-NL	AMBER-TI	BLEU	F15	F15G3	ROSE	TER	METEOR-1.3-RANK	MP4IBM1	MPF	TESLA-B	TESLA-F	TESLA-M	WMPF
English-French News Task															
CMU-HEAFIELD-COMBO	0.25	0.25	0.26	0.34	0.35	0.23	2.02	0.5	0.57	–41	52	0.54	–0.01	0.60	50
CU-ZEMAN	0.18	0.17	0.18	0.13	0.19	0.09	1.96	0.68	0.39	–46	35	0.34	–0.03	0.40	33
JHU	0.23	0.23	0.24	0.27	0.31	0.19	2.01	0.53	0.52	–43	47	0.49	–0.01	0.55	45
KIT	0.24	0.23	0.24	0.29	0.31	0.19	2.01	0.52	0.53	–42	49	0.51	–0.01	0.57	47
LATL-GENEVA	0.20	0.2	0.21	0.19	0.23	0.12	1.99	0.62	0.44	–43	41	0.44	–0.02	0.51	39
LIMSI	0.24	0.24	0.24	0.3	0.31	0.19	2.01	0.53	0.53	–41	49	0.51	–0.01	0.58	48
LIUM	0.24	0.23	0.24	0.29	0.31	0.19	2.01	0.53	0.53	–42	49	0.51	–0.01	0.57	47
ONLINE-A	0.24	0.23	0.24	0.27	0.3	0.18	2.01	0.53	0.52	–42	47	0.5	–0.01	0.56	46
ONLINE-B	0.25	0.25	0.25	0.33	0.35	0.23	2.02	0.5	0.56	–42	51	0.53	–0.01	0.59	50
RBMT-1	0.23	0.22	0.23	0.24	0.27	0.16	2.00	0.56	0.5	–41	45	0.48	–0.02	0.56	44
RBMT-2	0.22	0.21	0.22	0.22	0.25	0.14	1.99	0.58	0.47	–42	44	0.46	–0.02	0.53	42
RBMT-3	0.23	0.22	0.23	0.25	0.28	0.16	2.00	0.56	0.5	–41	46	0.48	–0.02	0.56	44
RBMT-4	0.22	0.21	0.22	0.23	0.26	0.15	1.99	0.58	0.47	–42	43	0.45	–0.02	0.51	42
RBMT-5	0.22	0.22	0.23	0.23	0.27	0.15	2	0.57	0.49	–41	45	0.47	–0.02	0.55	43
RWTH-HUCK	0.23	0.23	0.24	0.29	0.30	0.18	2.01	0.54	0.52	–42	48	0.5	–0.01	0.56	47
UEDIN	0.23	0.22	0.23	0.27	0.3	0.18	2.01	0.54	0.51	–42	47	0.49	–0.01	0.55	46
UPPSALA	0.23	0.22	0.23	0.27	0.29	0.17	2.00	0.55	0.51	–42	46	0.48	–0.01	0.55	45
UPPSALA-FBK	0.23	0.23	0.23	0.28	0.29	0.18	2.01	0.55	0.51	–42	47	0.49	–0.01	0.55	46
UPV-PRHLT-COMBO	0.25	0.24	0.25	0.32	0.34	0.22	2.02	0.50	0.55	–41	51	0.53	–0.01	0.59	49

Table 45: Automatic evaluation metric scores for systems in the WMT11 English-French News Task (newssyscombtest2011)

	AMBER	AMBER-NL	AMBER-TI	BLEU	F15	F15G3	ROSE	TER	METEOR-1.3-RANK	MP4IBM1	MPF	TESLA-B	TESLA-F	TESLA-M	WMPF
English-Spanish News Task															
CEU-UPV	0.24	0.24	0.24	0.29	0.3	0.18	2.01	0.51	0.55	–45	46	0.45	0.01	0.45	45
CMU-HEAFIELD-COMBO	0.26	0.25	0.26	0.35	0.34	0.22	2.02	0.47	0.58	–44	50	0.49	0.01	0.49	49
CU-ZEMAN	0.23	0.22	0.23	0.22	0.27	0.15	1.99	0.55	0.52	–48	39	0.41	0.00	0.41	38
KOC	0.23	0.23	0.23	0.25	0.27	0.16	2	0.54	0.52	–46	43	0.42	0.00	0.43	42
KOC-COMBO	0.25	0.24	0.25	0.31	0.32	0.2	2.01	0.5	0.56	–44	47	0.46	0.01	0.47	46
ONLINE-A	0.25	0.24	0.25	0.31	0.32	0.2	2.01	0.49	0.56	–44	48	0.46	0.01	0.47	46
ONLINE-B	0.25	0.25	0.25	0.33	0.32	0.2	2.02	0.50	0.57	–44	49	0.47	0.01	0.47	48
PROMT	0.24	0.23	0.24	0.28	0.28	0.17	2.00	0.53	0.52	–45	45	0.44	0.01	0.46	43
RBMT-1	0.23	0.23	0.23	0.25	0.27	0.16	2	0.55	0.51	–45	43	0.42	0.00	0.44	42
RBMT-2	0.23	0.22	0.23	0.25	0.26	0.15	1.99	0.55	0.5	–44	43	0.41	0.00	0.42	41
RBMT-3	0.24	0.23	0.24	0.28	0.28	0.17	2.00	0.53	0.52	–44	45	0.43	0.00	0.45	43
RBMT-4	0.23	0.22	0.23	0.26	0.26	0.16	1.99	0.54	0.51	–44	44	0.42	0.00	0.43	42
RBMT-5	0.23	0.22	0.23	0.24	0.26	0.15	1.99	0.57	0.49	–45	42	0.41	0.00	0.43	41
UEDIN	0.24	0.24	0.24	0.31	0.3	0.18	2.01	0.51	0.55	–45	47	0.45	0.01	0.45	46
UOW	0.23	0.23	0.24	0.28	0.28	0.16	2.00	0.53	0.53	–45	45	0.42	0.01	0.43	44
UOW-COMBO	0.25	0.25	0.25	0.33	0.32	0.2	2.01	0.50	0.56	–44	49	0.47	0.01	0.47	47
UPM	0.21	0.21	0.21	0.21	0.22	0.12	1.98	0.61	0.47	–47	39	0.37	0.00	0.37	38
UPPSALA	0.24	0.24	0.24	0.3	0.29	0.18	2.01	0.51	0.54	–45	46	0.44	0.01	0.44	45
UPV-PRHLT-COMBO	0.25	0.25	0.25	0.33	0.32	0.21	2.02	0.49	0.57	–44	49	0.47	0.01	0.48	48

Table 46: Automatic evaluation metric scores for systems in the WMT11 English-Spanish News Task (newssyscombtest2011)

	BLEU	MTERATER	MTERATER-PLUS	ROSE	TER	METEOR-1.3-ADQ	METEOR-1.3-RANK	MPF	TESLA-B	TESLA-F	TESLA-M	WMPF
Haitian Creole (clean)-English Haitian Creole SMS Emergency Response Featured Translation Task												
BM-I2R	0.33	–6798	–4575	1.96	0.51	0.62	0.34	43	0.44	0.03	0.46	43
CMU-DENKOWSKI	0.29	–6849	–6172	1.95	0.53	0.58	0.32	40	0.39	0.02	0.40	39
CMU-HEAFIELD-COMBO	0.32	–6188	–4347	1.96	0.51	0.61	0.34	42	0.43	0.03	0.45	42
CMU-HEWAVITHARANA	0.28	–6523	–6341	1.95	0.57	0.57	0.32	39	0.38	0.02	0.40	38
HYDERABAD	0.14	–7548	–8502	1.92	0.66	0.50	0.28	26	0.3	0.02	0.30	26
KOC	0.23	–6490	–9020	1.94	0.67	0.49	0.27	36	0.32	0.02	0.34	35
KOC-COMBO	0.29	–4901	–5349	1.95	0.57	0.56	0.31	39	0.38	0.02	0.4	39
LIU	0.27	–6526	–6078	1.95	0.59	0.56	0.31	38	0.38	0.02	0.39	37
UMD-EIDELMAN	0.26	–4407	–6215	1.95	0.57	0.55	0.31	38	0.37	0.02	0.4	37
UMD-HU	0.22	–6379	–7460	1.94	0.59	0.51	0.28	35	0.36	0.02	0.39	34
UPPSALA	0.27	–5497	–6754	1.95	0.59	0.54	0.3	38	0.36	0.02	0.39	37
UPV-PRHLT-COMBO	0.32	–6896	–5968	1.96	0.53	0.6	0.33	42	0.41	0.02	0.43	41

Table 47: Automatic evaluation metric scores for systems in the WMT11 Haitian Creole (clean)-English Haitian Creole SMS Emergency Response Featured Translation Task (newssyscombtest2011)

	BLEU	MTERATER	MTERATER-PLUS	ROSE	TER	METEOR-1.3-ADQ	METEOR-1.3-RANK	MPF	TESLA-B	TESLA-F	TESLA-M	WMPF
Haitian Creole (raw)-English Haitian Creole SMS Emergency Response Featured Translation Task												
BM-I2R	0.29	–3885	–3017	1.96	0.57	0.57	0.32	39	0.42	0.02	0.44	38
CMU-DENKOWSKI	0.25	–3965	–3905	1.95	0.60	0.53	0.3	35	0.38	0.02	0.4	35
CMU-HEAFIELD-COMBO	0.28	–3057	–2588	1.96	0.57	0.57	0.32	39	0.42	0.02	0.44	38
CMU-HEWAVITHARANA	0.25	–3701	–3824	1.95	0.61	0.53	0.3	35	0.37	0.02	0.39	35
JHU	0.14	–3207	–4279	1.92	0.74	0.43	0.24	26	0.30	0.02	0.32	26
LIU	0.25	–3447	–3445	1.95	0.60	0.54	0.30	36	0.38	0.02	0.4	35
UMD-EIDELMAN	0.24	–2826	–3754	1.94	0.64	0.52	0.29	34	0.36	0.02	0.39	34
UPV-PRHLT-COMBO	0.28	–3591	–3370	1.95	0.58	0.56	0.32	38	0.4	0.02	0.42	38

Table 48: Automatic evaluation metric scores for systems in the WMT11 Haitian Creole (raw)-English Haitian Creole SMS Emergency Response Featured Translation Task (newssyscombtest2011)

INTER-ANNOTATOR AGREEMENT (I.E. ACROSS ANNOTATORS)

	ALL COMPARISONS			NO REF COMPARISONS		
	$P(A)$	$P(E)$	κ	$P(A)$	$P(E)$	κ
Czech-English, individual systems	0.591	0.354	0.367	0.535	0.343	0.293
English-Czech, individual systems	0.608	0.359	0.388	0.552	0.350	0.312
German-English, individual systems	0.562	0.377	0.298	0.536	0.370	0.264
English-German, individual systems	0.564	0.352	0.327	0.528	0.348	0.276
Spanish-English, individual systems	0.695	0.398	0.493	0.683	0.393	0.477
English-Spanish, individual systems	0.574	0.343	0.352	0.548	0.339	0.317
French-English, individual systems	0.616	0.367	0.393	0.584	0.361	0.349
English-French, individual systems	0.631	0.382	0.403	0.603	0.376	0.363
European languages, individual systems	0.601	0.362	0.375	0.561	0.355	0.320
Czech-English, system combinations	0.700	0.334	0.549	0.577	0.369	0.329
English-Czech, system combinations	0.812	0.348	0.711	0.696	0.392	0.500
German-English, system combinations	0.675	0.353	0.498	0.629	0.341	0.437
English-German, system combinations	0.608	0.346	0.401	0.547	0.334	0.320
Spanish-English, system combinations	0.638	0.335	0.456	0.604	0.359	0.382
English-Spanish, system combinations	0.657	0.335	0.485	0.603	0.371	0.369
French-English, system combinations	0.654	0.336	0.479	0.608	0.336	0.410
English-French, system combinations	0.678	0.352	0.503	0.595	0.339	0.388
European languages, system combinations	0.671	0.335	0.505	0.598	0.342	0.389
Haitian (Clean)-English, individual systems	0.693	0.364	0.517	0.640	0.353	0.443
Haitian (Raw)-English, individual systems	0.689	0.357	0.517	0.639	0.344	0.450
Haitian-English, individual systems	0.691	0.362	0.516	0.639	0.350	0.446
Haitian (Clean)-English, system combinations	0.770	0.367	0.636	0.645	0.333	0.468
Haitian (Raw)-English, system combinations	0.745	0.345	0.611	0.753	0.361	0.613
Haitian-English, system combinations	0.761	0.358	0.628	0.674	0.335	0.509
Tunable metrics task (Urdu-English)	0.692	0.337	0.535	0.641	0.363	0.437
WMT10 (European languages, individual vs. individual)	0.663	0.394	0.445	0.620	0.385	0.382
WMT10 (European languages, combo vs. combo)	0.728	0.344	0.586	0.629	0.334	0.443
WMT10 (European languages, individual vs. combo)	N/A	N/A	N/A	0.634	0.360	0.428
WMT10 (European languages, all systems)	0.658	0.374	0.454	0.626	0.367	0.409

Table 49: Inter-annotator agreement rates, for the various manual evaluation tracks of WMT11, broken down by language pair. The highlighted rows correspond to rows in the top half of Table 7. See Table 50 below for detailed *intra*-annotator agreement rates.

INTRA-ANNOTATOR AGREEMENT (I.E. SELF-CONSISTENCY)

	ALL COMPARISONS			NO REF COMPARISONS		
	$P(A)$	$P(E)$	κ	$P(A)$	$P(E)$	κ
Czech-English, individual systems	0.762	0.354	0.632	0.713	0.343	0.564
English-Czech, individual systems	0.743	0.359	0.598	0.700	0.350	0.539
German-English, individual systems	0.675	0.377	0.478	0.670	0.370	0.475
English-German, individual systems	0.704	0.352	0.543	0.700	0.348	0.541
Spanish-English, individual systems	0.750	0.398	0.585	0.719	0.393	0.537
English-Spanish, individual systems	0.644	0.343	0.458	0.601	0.339	0.396
French-English, individual systems	0.829	0.367	0.730	0.816	0.361	0.712
English-French, individual systems	0.716	0.382	0.541	0.681	0.376	0.488
European languages, individual systems	0.722	0.362	0.564	0.685	0.355	0.512
Czech-English, system combinations	0.756	0.334	0.633	0.657	0.369	0.457
English-Czech, system combinations	0.923	0.348	0.882	0.842	0.392	0.740
German-English, system combinations	0.732	0.353	0.586	0.716	0.341	0.569
English-German, system combinations	0.722	0.346	0.575	0.676	0.334	0.513
Spanish-English, system combinations	0.783	0.335	0.673	0.720	0.359	0.562
English-Spanish, system combinations	0.741	0.335	0.610	0.711	0.371	0.540
French-English, system combinations	0.772	0.336	0.657	0.659	0.336	0.487
English-French, system combinations	0.841	0.352	0.755	0.714	0.339	0.568
European languages, system combinations	0.787	0.335	0.680	0.717	0.342	0.571
Haitian (Clean)-English, individual systems	0.758	0.364	0.619	0.686	0.353	0.515
Haitian (Raw)-English, individual systems	0.783	0.357	0.663	0.756	0.344	0.628
Haitian-English, individual systems	0.763	0.362	0.628	0.700	0.350	0.539
Haitian (Clean)-English, system combinations	0.882	0.367	0.813	0.778	0.333	0.667
Haitian (Raw)-English, system combinations	0.882	0.345	0.820	0.802	0.361	0.690
Haitian-English, system combinations	0.882	0.358	0.816	0.784	0.335	0.675
Tunable metrics task (Urdu-English)	0.857	0.337	0.784	0.856	0.363	0.774
WMT10 (European languages, individual vs. individual)	0.757	0.394	0.599	0.728	0.385	0.557
WMT10 (European languages, combo vs. combo)	0.783	0.344	0.670	0.719	0.334	0.578
WMT10 (European languages, individual vs. combo)	N/A	N/A	N/A	0.746	0.360	0.603
WMT10 (European languages, all systems)	0.755	0.374	0.609	0.734	0.367	0.580

Table 50: Intra-annotator agreement rates, for the various manual evaluation tracks of WMT11, broken down by language pair. The highlighted rows correspond to rows in the bottom half of Table 7. See Table 49 above for detailed *inter*-annotator agreement rates.