

# Further Meta-Evaluation of Machine Translation

**Chris Callison-Burch**  
Johns Hopkins University  
ccb@cs.jhu.edu

**Cameron Fordyce**  
camfordyce@gmail.com

**Philipp Koehn**  
University of Edinburgh  
pkoehn@inf.ed.ac.uk

**Christof Monz**  
Queen Mary, University of London  
christof@dcs.qmul.ac.uk

**Josh Schroeder**  
University of Edinburgh  
j.schroeder@ed.ac.uk

## Abstract

This paper analyzes the translation quality of machine translation systems for 10 language pairs translating between Czech, English, French, German, Hungarian, and Spanish. We report the translation quality of over 30 diverse translation systems based on a large-scale manual evaluation involving hundreds of hours of effort. We use the human judgments of the systems to analyze automatic evaluation metrics for translation quality, and we report the strength of the correlation with human judgments at both the system-level and at the sentence-level. We validate our manual evaluation methodology by measuring intra- and inter-annotator agreement, and collecting timing information.

## 1 Introduction

This paper<sup>1</sup> presents the results the shared tasks of the 2008 ACL Workshop on Statistical Machine Translation, which builds on two past workshops (Koehn and Monz, 2006; Callison-Burch et al., 2007). There were two shared tasks this year: a translation task which evaluated translation between 10 pairs of European languages, and an evaluation task which examines automatic evaluation metrics.

<sup>1</sup>This paper was corrected subsequent to publication. Tables 8 and 9, which summarize the strength of correlation of the automatic evaluation metrics, incorrectly averaged the correlations reported in Appendix B. This error resulted in the automatic evaluation metrics being misranked. We have corrected these tables as well as the text that describes them in Sections 6 and 8. This corrected version was released on August 30, 2008.

There were a number of differences between this year’s workshop and last year’s workshop:

- **Test set selection** – Instead of creating our test set by reserving a portion of the training data, we instead hired translators to translate a set of newspaper articles from a number of different sources. This out-of-domain test set contrasts with the in-domain Europarl test set.
- **New language pairs** – We evaluated the quality of Hungarian-English machine translation. Hungarian is a challenging language because it is agglutinative, has many cases and verb conjugations, and has freer word order. German-Spanish was our first language pair that did not include English, but was not manually evaluated since it attracted minimal participation.
- **System combination** – Saarland University entered a system combination over a number of rule-based MT systems, and provided their output, which were also treated as fully fledged entries in the manual evaluation. Three additional groups were invited to apply their system combination algorithms to all systems.
- **Refined manual evaluation** – Because last year’s study indicated that fluency and adequacy judgments were slow and unreliable, we dropped them from manual evaluation. We replaced them with yes/no judgments about the acceptability of translations of shorter phrases.
- **Sentence-level correlation** – In addition to measuring the correlation of automatic evalu-

ation metrics with human judgments at the system level, we also measured how consistent they were with the human rankings of individual sentences.

The remainder of this paper is organized as follows: Section 2 gives an overview of the shared translation task, describing the test sets, the materials that were provided to participants, and a list of the groups who participated. Section 3 describes the manual evaluation of the translations, including information about the different types of judgments that were solicited and how much data was collected. Section 4 presents the results of the manual evaluation. Section 5 gives an overview of the shared evaluation task, describes which automatic metrics were submitted, and tells how they were evaluated. Section 6 presents the results of the evaluation task. Section 7 validates the manual evaluation methodology.

## 2 Overview of the shared translation task

The shared translation task consisted of 10 language pairs: English to German, German to English, English to Spanish, Spanish to English, English to French, French to English, English to Czech, Czech to English, Hungarian to English, and German to Spanish. Each language pair had two test sets drawn from the proceedings of the European parliament, or from newspaper articles.<sup>2</sup>

### 2.1 Test data

The test data for this year’s task differed from previous years’ data. Instead of only reserving a portion of the training data as the test set, we hired people to translate news articles that were drawn from a variety of sources during November and December of 2007. We refer to this as the News test set. A total of 90 articles were selected, 15 each from a variety of Czech-, English-, French-, German-, Hungarian- and Spanish-language news sites.<sup>3</sup>

**Hungarian:** Napi (3 documents), Index (2),

<sup>2</sup>For Czech news editorials replaced the European parliament transcripts as the second test set, and for Hungarian the newspaper articles was the only test set.

<sup>3</sup>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.

Original source language	avg. BLEU
Hungarian	8.8
German	11.0
Czech	15.2
Spanish	17.3
English	17.7
French	18.6

Table 1: Difficulty of the test set parts based on the original language. For each part, we average BLEU scores from the Edinburgh systems for 12 language pairs of the shared task.

Origo (5), Népszabadság (2), HVG (2), Uniospez (1)

**Czech:** Aktuálně (1), iHNed (4), Lidovky (7), Novinky (3)

**French:** Liberation (4), Le Figaro (4), Dernieres Nouvelles (2), Les Echos (3), Canoe (2)

**Spanish:** Cinco Dias (7), ABC.es (3), El Mundo (5)

**English:** BBC (3), Scotsman (3), Economist (3), Times (3), New York Times (3)

**German:** Financial Times Deutschland (3), Süddeutsche Zeitung (3), Welt (3), Frankfurter Allgemeine Zeitung (3), Spiegel (3)

The translations were created by the members of EuroMatrix consortium who hired a mix of professional and non-professional translators. All translators were fluent or native speakers of both languages, and all translations were proofread by a native speaker of the target language. All of the translations were done directly, and not via an intermediate language. So for instance, each of the 15 Hungarian articles were translated into Czech, English, French, German and Spanish. The total cost of creating the 6 test sets consisting of 2,051 sentences in each language was approximately 17,200 euros (around 26,500 dollars at current exchange rates, at slightly more than 10c/word).

Having a test set that is balanced in six different source languages and translated across six languages raises some interesting questions. For instance, is it easier, when the machine translation system translates in the same direction as the human translator? We found no conclusive evidence that shows this. What is striking, however, that the parts

differ dramatically in difficulty, based on the original source language. For instance the Edinburgh French-English system has a BLEU score of 26.8 on the part that was originally Spanish, but a score of 9.7 on the part that was originally Hungarian. For average scores for each original language, see Table 1.

In order to remain consistent with previous evaluations, we also created a Europarl test set. The Europarl test data was again drawn from the transcripts of EU parliamentary proceedings from the fourth quarter of 2000, which is excluded from the Europarl training data. Our rationale behind investing a considerable sum to create the News test set was that we believe that it more accurately represents the quality of systems' translations than when we simply hold out a portion of the training data as the test set, as with the Europarl set. For instance, statistical systems are heavily optimized to their training data, and do not perform as well on out-of-domain data (Koehn and Schroeder, 2007). Having both the News test set and the Europarl test set allows us to contrast the performance of systems on in-domain and out-of-domain data, and provides a fairer comparison between systems trained on the Europarl corpus and systems that were developed without it.

## 2.2 Provided materials

To lower the barrier of entry for newcomers to the field, we provided a complete baseline MT system, along with data resources. We provided:

- sentence-aligned training corpora
- language model data
- development and dev-test sets
- Moses open source toolkit for phrase-based statistical translation (Koehn et al., 2007)

The performance of this baseline system is similar to the best submissions in last year's shared task.

The training materials are described in Figure 1.

## 2.3 Submitted systems

We received submissions from 23 groups from 18 institutions, as listed in Table 2. We also evaluated seven additional commercial rule-based MT systems, bringing the total to 30 systems. This is

a significant increase over last year's shared task, where there were submissions from 15 groups from 14 institutions. Of the 15 groups that participated in last year's shared task, 11 groups returned this year. One of the goals of the workshop was to attract submissions from newcomers to the field, and we are pleased to have attracted many smaller groups, some as small as a single graduate student and her advisor.

The 30 submitted systems represent a broad range of approaches to statistical machine translation. These include statistical phrase-based and rule-based (RBMT) systems (which together made up the bulk of the entries), and also hybrid machine translation, and statistical tree-based systems. For most language pairs, we assembled a solid representation of the state of the art in machine translation.

In addition to individual systems being entered, this year we also solicited a number of entries which combined the results of other systems. We invited researchers at BBN, Carnegie Mellon University, and the University of Edinburgh to apply their system combination algorithms to all of the systems submitted to shared translation task. We designated the translations of the Europarl set as the development data for combination techniques which weight each system.<sup>4</sup> CMU combined the French-English systems, BBN combined the French-English and German-English systems, and Edinburgh submitted combinations for the French-English and German-English systems as well as a multi-source system combination which combined all systems which translated from any language pair into English for the News test set. The University of Saarland also produced a system combination over six commercial RBMT systems (Eisele et al., 2008). Saarland graciously provided the output of these systems, which we manually evaluated alongside all other entries.

For more on the participating systems, please refer to the respective system descriptions in the proceedings of the workshop.

---

<sup>4</sup>Since the performance of systems varied significantly between the Europarl and News test sets, such weighting might not be optimal. However this was a level playing field, since none of the individual systems had development data for the News set either.

## Europarl Training Corpus

	Spanish ↔ English		French ↔ English		German ↔ English		German ↔ Spanish	
<b>Sentences</b>	1,258,778		1,288,074		1,266,520		1,237,537	
<b>Words</b>	36,424,186	35,060,653	38,784,144	36,046,219	33,404,503	35,259,758	32,652,649	35,780,165
<b>Distinct words</b>	149,159	96,746	119,437	97,571	301,006	96,802	298,040	148,206

## News Commentary Training Corpus

	Spanish ↔ English		French ↔ English		German ↔ English		German ↔ Spanish	
<b>Sentences</b>	64,308		55,030		72,291		63,312	
<b>Words</b>	1,759,972	1,544,633	1,528,159	1,329,940	1,784,456	1,718,561	1,597,152	1,751,215
<b>Distinct words</b>	52,832	38,787	42,385	36,032	84,700	40,553	78,658	52,397

## Hunglish Training Corpus

	Hungarian ↔ English	
<b>Sentences</b>	1,517,584	
<b>Words</b>	26,082,667	31,458,540
<b>Distinct words</b>	717,198	192,901

## CzEng Training Corpus

	Czech ↔ English	
<b>Sentences</b>	1,096,940	
<b>Words</b>	15,336,783	17,909,979
<b>Distinct words</b>	339,683	129,176

## Europarl Language Model Data

	English	Spanish	French	German
<b>Sentence</b>	1,412,546	1,426,427	1,438,435	1,467,291
<b>Words</b>	34,501,453	36,147,902	35,680,827	32,069,151
<b>Distinct words</b>	100,826	155,579	124,149	314,990

## Europarl test set

	English	Spanish	French	German
<b>Sentences</b>	2,000			
<b>Words</b>	60,185	61,790	64,378	56,624
<b>Distinct words</b>	6,050	7,814	7,361	8,844

## News Commentary test set

	English	Czech
<b>Sentences</b>	2,028	
<b>Words</b>	45,520	39,384
<b>Distinct words</b>	7,163	12,570

## News Test Set

	English	Spanish	French	German	Czech	Hungarian
<b>Sentences</b>	2,051					
<b>Words</b>	43,482	47,155	46,183	41,175	36,359	35,513
<b>Distinct words</b>	7,807	8,973	8,898	10,569	12,732	13,144

Figure 1: Properties of the training and test sets used in the shared task. The training data is drawn from the Europarl corpus and from the Project Syndicate, a web site which collects political commentary in multiple languages. For Czech and Hungarian we use other available parallel corpora. Note that the number of words is computed based on the provided tokenizer and that the number of distinct words is the based on lowercased tokens.

<b>ID</b>	<b>Participant</b>
BBN-COMBO	BBN system combination (Rosti et al., 2008)
CMU-COMBO	Carnegie Mellon University system combination (Jayaraman and Lavie, 2005)
CMU-GIMPEL	Carnegie Mellon University Gimpel (Gimpel and Smith, 2008)
CMU-SMT	Carnegie Mellon University SMT (Bach et al., 2008)
CMU-STATXFER	Carnegie Mellon University Stat-XFER (Hanneman et al., 2008)
CU-TECTOMT	Charles University TectoMT (Zabokrtsky et al., 2008)
CU-BOJAR	Charles University Bojar (Bojar and Hajič, 2008)
CUED	Cambridge University (Blackwood et al., 2008)
DCU	Dublin City University (Tinsley et al., 2008)
LIMSI	LIMSI (Déchelotte et al., 2008)
LIU	Linköping University (Stymne et al., 2008)
LIUM-SYSTRAN	LIUM / Systran (Schwenk et al., 2008)
MLOGIC	Morphologic (Novák et al., 2008)
PCT	a commercial MT provider from the Czech Republic
RBMT1–6	Babelfish, Lingenio, Lucy, OpenLogos, ProMT, SDL (ordering anonymized)
SAAR	University of Saarbruecken (Eisele et al., 2008)
SYSTRAN	Systran (Dugast et al., 2008)
UCB	University of California at Berkeley (Nakov, 2008)
UCL	University College London (Wang and Shawe-Taylor, 2008)
UEDIN	University of Edinburgh (Koehn et al., 2008)
UEDIN-COMBO	University of Edinburgh system combination (Josh Schroeder)
UMD	University of Maryland (Dyer, 2007)
UPC	Universitat Politecnica de Catalunya, Barcelona (Khalilov et al., 2008)
UW	University of Washington (Axelrod et al., 2008)
XEROX	Xerox Research Centre Europe (Nikoulina and Dymetman, 2008)

Table 2: Participants in the shared translation task. Not all groups participated in all language pairs.

### 3 Human evaluation

As with last year’s workshop, 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, rather than select an official automatic evaluation metric like the NIST Machine Translation Workshop does (Przybocki and Peterson, 2008), 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 monumental effort to conduct it 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. More than 100 people participated in the manual evaluation, with 75 people putting in more than an hour’s worth of effort, and 25 putting in more than four hours. A collective total of 266 hours of labor was invested.

We wanted to ensure that we were using our annotators’ time effectively, so we carefully designed the manual evaluation process. In our analysis of last year’s manual evaluation we found that the NIST-style fluency and adequacy scores (LDC, 2005) were overly time consuming and inconsistent.<sup>5</sup> We therefore abandoned this method of evaluating the translations.

We asked people to evaluate the systems’ output in three different ways:

- Ranking translated sentences relative to each other
- Ranking the translations of syntactic constituents drawn from the source sentence
- Assigning absolute yes or no judgments to the translations of the syntactic constituents.

The manual evaluation software asked for repeated judgments from the same individual, and had multiple people judge the same item, and logged the

---

<sup>5</sup>It took 26 seconds on average to assign fluency and adequacy scores to a single sentence, and the inter-annotator agreement had a Kappa of between .225–.25, meaning that annotators assigned the same scores to identical sentences less than 40% of the time.

time it took to complete each judgment. This allowed us to measure intra- and inter-annotator agreement, and to analyze the average amount of time it takes to collect the different kinds of judgments. Our analysis is presented in Section 7.

#### 3.1 Ranking translations of sentences

Ranking translations relative to each other is a relatively intuitive and straightforward task. We therefore kept the instructions simple. The instructions for this task were:

Rank each whole sentence translation from Best to Worst relative to the other choices (ties are allowed).

Ranking several translations at a time is a variant of force choice judgments where a pair of systems is presented and an annotator is asked “Is A better than B, worse than B, or equal to B.” In our experiments, annotators were shown five translations at a time, except for the Hungarian and Czech language pairs where there were fewer than five system submissions. In most cases there were more than 5 systems submissions. We did not attempt to get a complete ordering over the systems, and instead relied on random selection and a reasonably large sample size to make the comparisons fair.

#### 3.2 Ranking translations of syntactic constituents

We continued the constituent-based evaluation that we piloted last year, wherein we solicited judgments about the translations of short phrases within sentences rather than whole sentences. We parsed the source language sentence, selected syntactic constituents from the tree, and had people judge the translations of those syntactic phrases. In order to draw judges’ attention to these regions, we highlighted the selected source phrases and the corresponding phrases in the translations. The corresponding phrases in the translations were located via automatic word alignments.

Figure 2 illustrates how the source and reference phrases are highlighted via automatic word alignments. The same is done for sentence and each of the system translations. The English, French, German and Spanish test sets were automatically

Language Pair	Test Set	Constituent Rank	Yes/No Judgments	Sentence Ranking
English-German	Europarl	2,032	2,034	1,004
	News	2,170	2,221	1,115
German-English	Europarl	1,705	1,674	819
	News	1,938	1,881	1,944
English-Spanish	Europarl	1,200	1,247	615
	News	1,396	1,398	700
Spanish-English	Europarl	1,855	1,921	948
	News	2,063	1,939	1,896
English-French	Europarl	1,248	1,265	674
	News	1,741	1,734	843
French-English	Europarl	1,829	1,841	909
	News	2,467	2,500	2,671
English-Czech	News	2,069	2,070	1,045
	Commentary	1,840	1,815	932
Czech-English	News	0	0	1,400
	Commentary	0	0	1,731
Hungarian-English	News	0	0	937
All-English	News	0	0	4,868
<b>Totals</b>		<b>25,553</b>	<b>25,540</b>	<b>25,051</b>

Table 3: The number of items that were judged for each task during the manual evaluation. The All-English judgments were reused in the News task for individual language pairs.

parsed using high quality parsers for those languages (Bikel, 2002; Arun and Keller, 2005; Dubey, 2005; Bick, 2006).

The word alignments were created with Giza++ (Och and Ney, 2003) applied to a parallel corpus containing the complete Europarl training data, plus sets of 4,051 sentence pairs created by pairing the test sentences with the reference translations, and the test sentences paired with each of the system translations. The phrases in the translations were located using standard phrase extraction techniques (Koehn et al., 2003). Because the word-alignments were created automatically, and because the phrase extraction is heuristic, the phrases that were selected may not exactly correspond to the translations of the selected source phrase. We noted this in the instructions to judges:

Rank each constituent translation from Best to Worst relative to the other choices (ties are allowed). Grade **only the high-lighted part** of each translation.

*Please note that segments are selected au-*

*tomatically, and they should be taken as an approximate guide. They might include extra words that are not in the actual alignment, or miss words on either end.*

The criteria that we used to select which constituents to evaluate were:

- The constituent could not be the whole source sentence
- The constituent had to be longer three words, and be no longer than 15 words
- The constituent had to have a corresponding phrase with a consistent word alignment in each of the translations

The final criterion helped reduce the number of alignment errors, but may have biased the sample to phrases that are more easily aligned.

### 3.3 Yes/No judgments for the translations of syntactic constituents

This year we introduced a variant on the constituent-based evaluation, where instead of asking judges

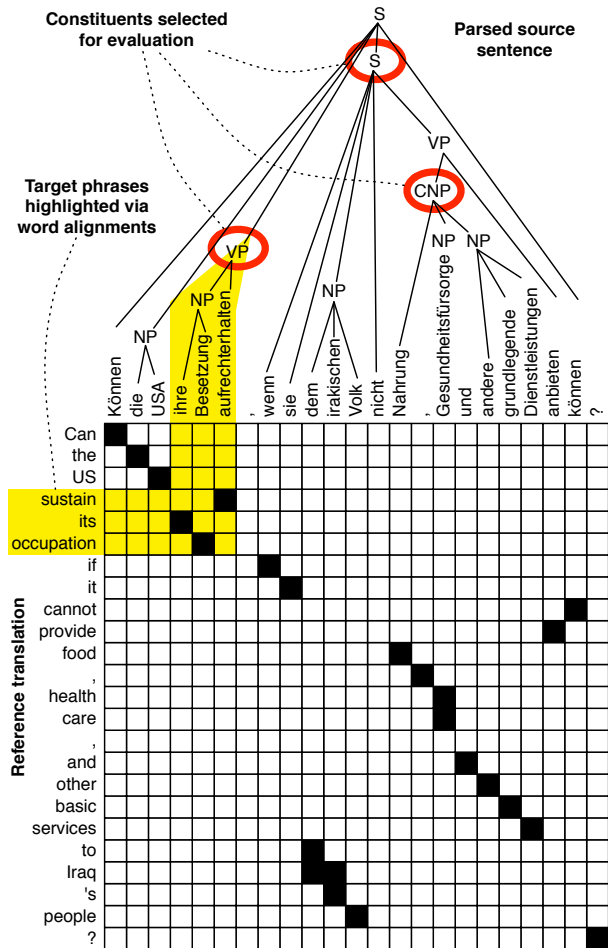


Figure 2: In constituent-based evaluation, the source sentence was parsed, and automatically aligned with the reference translation and systems’ translations

to rank the translations of phrases relative to each other, we asked them to indicate which phrasal translations were acceptable and which were not.

Decide if the **highlighted part** of each translation is acceptable, given the reference. This should not be a relative judgment against the other system translations.

The instructions also contained the same caveat about the automatic alignments as above. For each phrase the judges could click on “Yes”, “No”, or “Not Sure.” The number of times people clicked on “Not Sure” varied by language pair and task. It was selected as few as 5% of the time for the English-Spanish News task to as many as 12.5% for the Czech-English News task.

### 3.4 Collecting judgments

We collected judgments using a web-based tool that presented judges with batches of each type of evaluation. We presented them with five screens of sentence rankings, ten screens of constituent rankings, and ten screen of yes/no judgments. The order of the types of evaluation were randomized.

In order to measure intra-annotator agreement 10% of the items were repeated and evaluated twice by each judge. In order to measure inter-annotator agreement 40% of the items were randomly drawn from a common pool that was shared across all annotators so that we would have items that were judged by multiple annotators.

Judges were allowed to select whichever data set they wanted, and to evaluate translations into whatever languages they were proficient in. Shared task participants were excluded from judging their own systems.

In addition to evaluation each language pair individually, we also combined all system translations into English for the News test set, taking advantage of the fact that our test sets were parallel across all languages. This allowed us to gather interesting data about the difficulty of translating from different languages into English.

Table 3 gives a summary of the number of judgments that we collected for translations of individual sentences. We evaluated 14 translation tasks with three different types of judgments for most of them, for a total of 46 different conditions. In total we collected over 75,000 judgments. Despite the large number of conditions we managed to collect between 1,000–2,000 judgments for the constituent-based evaluation, and several hundred to several thousand judgments for the sentence ranking tasks.

## 4 Translation task results

Tables 4, 5, and 6 summarize the results of the human evaluation of the quality of the machine translation systems. Table 4 gives the results for the manual evaluation which ranked the translations of sentences. It shows the average number of times that systems were judged to be better than or equal to any other system. Table 5 similarly summarizes the results for the manual evaluation which ranked the translations of syntactic constituents. Table 6







UEDIN-COMBO <sub>xx</sub>	.717	SAAR <sub>fr</sub>	.584
LIUM-SYSTRAN-C <sub>fr</sub>	.708	SAAR-C <sub>de</sub>	.574
RBMT5 <sub>fr</sub>	.706	RBMT4 <sub>de</sub>	.573
UEDIN-COMBO <sub>fr</sub>	.704	CUED <sub>es</sub>	.572
LIUM-SYSTRAN <sub>fr</sub>	.702	RBMT3 <sub>de</sub>	.552
RBMT4 <sub>es</sub>	.699	CMU-SMT <sub>es</sub>	.548
LIMS1 <sub>fr</sub>	.699	UCB <sub>es</sub>	.547
BBN-COMBO <sub>fr</sub>	.695	LIMS1 <sub>es</sub>	.537
SAAR <sub>es</sub>	.678	RBMT6 <sub>de</sub>	.509
CUED-CONTRAST <sub>es</sub>	.674	RBMT5 <sub>de</sub>	.493
CMU-COMBO <sub>fr</sub>	.661	LIMS1 <sub>de</sub>	.469
UEDIN <sub>es</sub>	.654	LIU <sub>de</sub>	.447
CUED <sub>fr</sub>	.652	SAAR <sub>de</sub>	.445
CUED-CONTRAST <sub>fr</sub>	.638	CMU-STATXFR <sub>fr</sub>	.444
RBMT4 <sub>fr</sub>	.637	UMD <sub>cz</sub>	.429
UPC <sub>es</sub>	.633	BBN-COMBO <sub>de</sub>	.407
RBMT3 <sub>es</sub>	.628	UEDIN <sub>de</sub>	.402
RBMT2 <sub>de</sub>	.627	MORPHOLOGIC <sub>hu</sub>	.387
SAAR-CONTRAST <sub>fr</sub>	.624	DCU <sub>cz</sub>	.380
UEDIN <sub>fr</sub>	.616	UEDIN-COMBO <sub>de</sub>	.327
RBMT6 <sub>fr</sub>	.615	UEDIN <sub>cz</sub>	.293
RBMT6 <sub>es</sub>	.615	CMU-STATXFER <sub>de</sub>	.280
RBMT3 <sub>fr</sub>	.612	UEDIN <sub>hu</sub>	.188

Table 7: The average number of times that each system was judged to be better than or equal to all other systems in the sentence ranking task for the All-English condition. The subscript indicates the source language of the system.

for each language pair and test set, and found that RBMT was a surprisingly good approach in many of the conditions. RBMT generally did better on the News test set and for translations into German, suggesting that SMT’s forte is in test sets where it has appropriate tuning data and for language pairs with less reordering than between German and English.

## 5 Shared evaluation task overview

The manual evaluation data provides a rich source of information beyond simply analyzing the quality of translations produced by different systems. In particular, it is especially useful for validating the automatic metrics which are frequently used by the machine translation research community. We continued the shared task which we debuted last year, by examining how well various automatic metrics correlate with human judgments.

In addition to examining how well the automatic evaluation metrics predict human judgments at the system-level, this year we have also started to measure their ability to predict sentence-level judgments.

The automatic metrics that were evaluated in this year’s shared task were the following:

- Bleu (Papineni et al., 2002)—Bleu remains the *de facto* standard in machine translation evaluation. It calculates n-gram precision and a brevity penalty, and can make use of multiple reference translations as a way of capturing some of the allowable variation in translation. We use a single reference translation in our experiments.
- Meteor (Agarwal and Lavie, 2008)—Meteor measures precision and recall for unigrams and applies a fragmentation penalty. It uses flexible word matching based on stemming and WordNet-synonymy. A number of variants are investigated here: meteor-baseline and meteor-ranking are optimized for correlation with adequacy and ranking judgments respectively. mbleu and mter are Bleu and TER computed using the flexible matching used in Meteor.
- Gimenez and Marquez (2008) measure overlapping grammatical dependency relationships (DP), semantic roles (SR), and discourse representations (DR). The authors further investigate combining these with other metrics including TER, Bleu, GTM, Rouge, and Meteor (ULC and ULCh).
- Popovic and Ney (2007) automatically evaluate translation quality by examining sequences of parts of speech, rather than words. They calculate Bleu (posbleu) and F-measure (pos4gramFmeasure) by matching part of speech 4grams in a hypothesis translation against the reference translation.

In addition to the above metrics, which scored the translations on both the system-level<sup>6</sup> and the sentence-level, there were a number of metrics which focused on the sentence-level:

- Albrecht and Hwa (2008) use support vector regression to score translations using past WMT manual assessment data as training examples.

<sup>6</sup>We provide the scores assigned to each system by these metrics in Appendix A.

The metric uses features derived from target-side language models and machine-generated translations (svm-pseudo-ref) as well as reference human translations (svm-human-ref).

- Duh (2008) similarly used support vector machines to predict an ordering over a set of system translations (svm-rank). Features included in Duh (2008)’s training were sentence-level BLEU scores and intra-set ranks computed from the entire set of translations.
- USaar’s evaluation metric (alignment-prob) uses Giza++ to align outputs of multiple systems with the corresponding reference translations, with a bias towards identical one-to-one alignments through a suitably augmented corpus. The Model4 log probabilities in both directions are added and normalized to a scale between 0 and 1.

### 5.1 Measuring system-level correlation

To measure the correlation of the automatic metrics with the human judgments of translation quality at the system-level we used Spearman’s rank correlation coefficient  $\rho$ . We converted the raw scores assigned each system into ranks. We assigned a ranking to the systems for each of the three types of manual evaluation based on:

- The percent of time that the sentences it produced were judged to be better than or equal to the translations of any other system.
- The percent of time that its constituent translations were judged to be better than or equal to the translations of any other system.
- The percent of time that its constituent translations were judged to be acceptable.

We calculated  $\rho$  three times for each automatic metric, comparing it to each type of human evaluation. Since there were no ties  $\rho$  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 <sub>$i$</sub>  and  $n$  is the number of systems. The possible values of  $\rho$  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 value for  $\rho$  is making predictions that are more similar to the human judgments than an automatic evaluation metric with a lower  $\rho$ .

### 5.2 Measuring consistency at the sentence-level

Measuring sentence-level correlation under our human evaluation framework was made complicated by the fact that we abandoned the fluency and adequacy judgments which are intended to be absolute scales. Some previous work has focused on developing automatic metrics which predict human ranking at the sentence-level (Kulesza and Shieber, 2004; Albrecht and Hwa, 2007a; Albrecht and Hwa, 2007b). Such work generally used the 5-point fluency and adequacy scales to combine the translations of all sentences into a single ranked list. This list could be compared against the scores assigned by automatic metrics and used to calculate correlation coefficients. We did not gather any absolute scores and thus cannot compare translations across different sentences. Given the seemingly unreliable fluency and adequacy assignments that people make even for translations of the same sentences, it may be dubious to assume that their scoring will be reliable across sentences.

The data points that we have available consist of a set of 6,400 human judgments each ranking the output of 5 systems. It’s straightforward to construct a ranking of each of those 5 systems using the scores assigned to their translations of that sentence by the automatic evaluation metrics. When the automatic scores have been retrieved, we have 6,400 pairs of ranked lists containing 5 items. How best to treat these is an open discussion, and certainly warrants further thought. It does not seem like a good idea to calculate  $\rho$  for each pair of ranked list, because 5 items is an insufficient number to get a reliable correlation coefficient and its unclear if averaging over all 6,400 lists would make sense. Furthermore, many of the human judgments of 5 contained ties, further complicating matters.

Therefore rather than calculating a correlation coefficient at the sentence-level we instead ascertained how consistent the automatic metrics were with the human judgments. The way that we calculated con-

	RANK	CONST	YES/NO	OVERALL
DP	<b>.83</b>	.66	.74	.74
ULCh	<b>.83</b>	.68	.78	.76
DR	.8	.54	.64	.66
ULC	.8	.72	<b>.81</b>	<b>.77</b>
meteor-ranking	.78	.55	.62	.65
SR	.77	.66	.76	.73
posbleu	.75	.69	.78	.74
meteor-baseline	.75	.6	.63	.66
pos4gramFmeasure	.74	.59	.69	.67
bleu	.52	.72	.75	.66
mbleu	.5	<b>.75</b>	.71	.65
mter	.38	.73	.68	.6
svm-rank	.37	.1	.22	.23

Table 8: Average system-level correlations for the automatic evaluation metrics on translations into English (**This table was corrected after publication**)

sistency was the following: for every pairwise comparison of two systems on a single sentence by a person, we counted the automatic metric as being consistent if the relative scores were the same (i.e. the metric assigned a higher score to the higher ranked system). We divided this by the total number of pairwise comparisons to get a percentage. Because the systems generally assign real numbers as scores, we excluded pairs that the human annotators ranked as ties.

## 6 Evaluation task results

Tables 8 and 9 report the system-level  $\rho$  for each automatic evaluation metric, averaged over all translations directions into English and out of English.<sup>7</sup> For the into English direction Gimenez and Marquez’s method for measuring overlap in dependency relations (DP) had the strongest correlation with ranking the translations of whole sentences barely edging out their combination metric (ULCh). Bleu ranked 10th in the correlation with the human rankings based on whole sentences. Bleu’s poor performance may be partially attributable to the fact that we were using

<sup>7</sup>Table 9 excludes the English-Spanish News Task, since it had a negative correlation with most of the automatic metrics. See Tables 19 and 20.

	RANK	CONST	YES/NO	OVERALL
posbleu	<b>.75</b>	.78	.8	<b>.78</b>
pos4gramFmeasure	<b>.75</b>	<b>.81</b>	.79	<b>.78</b>
bleu	.68	.79	.79	.76
svm-rank	.67	.73	.8	.73
mbleu	.64	.8	<b>.81</b>	.75
meteor-baseline	.59	.78	.76	.71
meteor-ranking	.55	.74	.74	.68
mter	.54	.69	.73	.65

Table 9: Average system-level correlations for the automatic evaluation metrics on translations into French, German and Spanish (**This table was corrected after publication**)

test sets which had only a single reference translation, since the cost of creating multiple references was prohibitively expensive (see Section 2.1).

In the reverse direction, for translations out of English into the other languages, Bleu does considerably better, placing third overall and in its correlation with human rankings based on whole sentences. It is outperformed by two part-of-speech variants on it proposed by Popovic and Ney (2007). Appendix B gives a break down of the correlations for each of the language pairs and test sets.

Tables 10 and 11 report the consistency of the automatic evaluation metrics with human judgments on a sentence-by-sentence basis, rather than on the system level. For the translations into English Gimenez and Marquez’s ULC metric (which itself combines many other metrics) had the strongest correlation with human judgments, correctly predicting the human ranking of a each pair of system translations of a sentence more than half the time. This is dramatically higher than the chance baseline, which is not .5, since it must correctly rank a list of systems rather than a pair. For the reverse direction many metrics perform similarly.

## 7 Validation and analysis of the manual evaluation

In addition to scoring the shared task entries, we also continued on our campaign for improving the pro-

	RANK	CONST	YES/NO	OVERALL
ULC	<b>.56</b>	<b>.55</b>	<b>.56</b>	<b>.56</b>
ULCh	<b>.56</b>	.54	.54	.55
svm-human-ref	.54	.54	.55	.54
svm-pseudo-ref	.54	.54	.54	.54
alignment-prob	.52	.54	.54	.53
DP	.51	.53	.54	.53
meteor-baseline	.51	.52	.54	.52
meteor-ranking	.51	.52	.54	.52
mbleu	.51	.52	.54	.52
DR	.50	.51	.53	.51
SR	.50	.49	.51	.50
pos4gramFmeasure	.50	.52	.52	.51
svm-rank	.49	.50	.50	.50
mter	.44	.47	.48	.46
posbleu	.44	.43	.45	.44

Table 10: The percent of time that each automatic metric was consistent with human judgments for translations into English

cess of manual evaluation.

### 7.1 Inter- and Intra-annotator agreement

We measured pairwise agreement among annotators using the kappa coefficient ( $K$ ) which is widely used in computational linguistics for measuring agreement in category judgments (Carletta, 1996). It is defined as

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

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. We define chance agreement for ranking tasks as  $\frac{1}{3}$  since there are three possible outcomes when ranking the output of a pair of systems:  $A > B$ ,  $A = B$ ,  $A < B$ , and for the Yes/No judgments as  $\frac{1}{2}$  since we ignored those items marked “Not Sure”.

For inter-annotator agreement we calculated  $P(A)$  for the yes/no judgments by examining all items that were annotated by two or more annotators, and calculating the proportion of time they assigned identical scores to the same items. For the ranking tasks we calculated  $P(A)$  by examining all

	RANK	CONST	YES/NO	OVERALL
meteor-ranking	<b>.52</b>	.50	<b>.53</b>	<b>.52</b>
mbleu	<b>.52</b>	.52	.52	<b>.52</b>
pos4gramFmeasure	<b>.52</b>	<b>.53</b>	.51	<b>.52</b>
svm-rank	.51	<b>.53</b>	.51	<b>.52</b>
meteor-baseline	.51	.49	.52	.51
mter	.45	.44	.46	.45
posbleu	.44	.45	.42	.44

Table 11: The percent of time that each automatic metric was consistent with human judgments for translations into other languages

pairs of systems which had been judged by two or more judges, and calculated the proportion of time that they agreed that  $A > B$ ,  $A = B$ , or  $A < B$ . For intra-annotator agreement we did similarly, but gathered items that were annotated on multiple occasions by a single annotator.

Table 12 gives  $K$  values for inter-annotator agreement, and Table 13 gives  $K$  values for intra-annotator agreement. These give an indication of how often different judges agree, and how often single judges are consistent for repeated judgments, respectively. The interpretation of Kappa varies, but according to Landis and Koch (1977), 0–.2 is slight, .2–.4 is fair, .4–.6 is moderate, .6–.8 is substantial and the rest almost perfect. The inter-annotator agreement for the sentence ranking task was fair, for the constituent ranking it was moderate and for the yes/no judgments it was substantial.<sup>8</sup> For the intra-annotator agreement  $K$  indicated that people had moderate consistency with their previous judgments on the sentence ranking task, substantial consistency with their previous constituent ranking judgments, and nearly perfect consistency with their previous yes/no judgments.

These  $K$  values indicate that people are able to more reliably make simple yes/no judgments about the translations of short phrases than they are to

<sup>8</sup>Note that for the constituent-based evaluations we verified that the high  $K$  was not trivially due to identical phrasal translations. We excluded screens where all five phrasal translations presented to the annotator were identical, and report both numbers.

Evaluation type	$P(A)$	$P(E)$	$K$
Sentence ranking	.578	.333	.367
Constituent ranking	.671	.333	.506
Constituent (w/identicals)	.678	.333	.517
Yes/No judgments	.821	.5	.642
Yes/No (w/identicals)	.825	.5	.649

Table 12: Kappa coefficient values representing the inter-annotator agreement for the different types of manual evaluation

rank phrases or whole sentences. While this is an interesting observation, we do not recommend doing away with the sentence ranking judgments. The higher agreement on the constituent-based evaluation may be influenced based on the selection criteria for which phrases were selected for evaluation (see Section 3.2). Additionally, the judgments of the short phrases are not a great substitute for sentence-level rankings, at least in the way we collected them. The average correlation coefficient between the constituent-based judgments with the sentence ranking judgments is only  $\rho = .51$ . Tables 19 and 20 give a detailed break down of the correlation of the different types of human judgments with each other on each translation task. It may be possible to select phrases in such a way that the constituent-based evaluations are a better substitute for the sentence-based ranking, for instance by selecting more of constituents from each sentence, or attempting to cover most of the words in each sentence in a phrase-by-phrase manner. This warrants further investigation. It might also be worthwhile to refine the instructions given to annotators about how to rank the translations of sentences to try to improve their agreement, which is currently lower than we would like it to be (although it is substantially better than the previous fluency and adequacy scores, which had a  $K < .25$  in last year’s evaluation).

## 7.2 Timing

We used the web interface to collect timing information. The server recorded the time when a set of sentences was given to a judge and the time when the judge returned the sentences. It took annotators an average of 18 seconds per sentence to rank a list

Evaluation type	$P(A)$	$P(E)$	$K$
Sentence ranking	.691	.333	.537
Constituent ranking	.825	.333	.737
Constituent (w/identicals)	.832	.333	.748
Yes/No judgments	.928	.5	.855
Yes/No (w/identicals)	.930	.5	.861

Table 13: Kappa coefficient values for intra-annotator agreement for the different types of manual evaluation

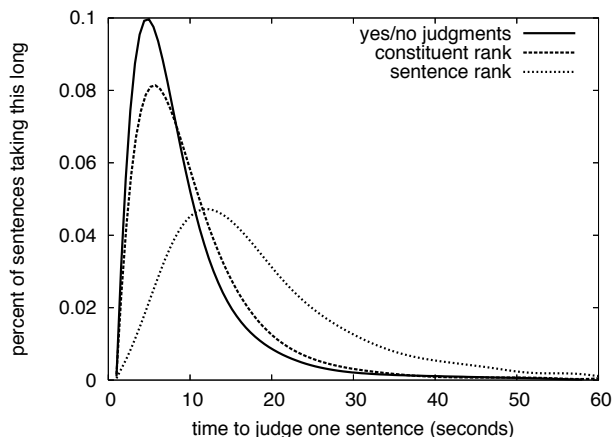


Figure 3: Distributions of the amount of time it took to judge single sentences for the three types of manual evaluation

of sentences.<sup>9</sup> It took an average of 10 seconds per sentence for them to rank constituents, and an average of 8.5 seconds per sentence for them to make yes/no judgments. Figure 3 shows the distribution of times for these tasks.

These timing figures indicate that the tasks which the annotators were the most reliable on (yes/no judgments and constituent ranking) were also much quicker to complete than the ones they were less reliable on (ranking sentences). Given that they are faster at judging short phrases, they can do proportionally more of them. For instance, we could collect 211 yes/no judgments in the same amount of time that it would take us to collect 100 sentence ranking judgments. However, this is partially offset by the fact that many of the translations of shorter phrases are identical, which means that we have to collect

<sup>9</sup>Sets which took longer than 5 minutes were excluded from these calculations, because there was a strong chance that annotators were interrupted while completing the task.

more judgments in order to distinguish between two systems.

### 7.3 The potential for re-usability of human judgments

One strong advantage of the yes/no judgments over the ranking judgments is their potential for reuse. We have invested hundreds of hours worth of effort evaluating the output of the translation systems submitted to this year's workshop and last year's workshop. While the judgments that we collected provide a wealth of information for developing automatic evaluation metrics, we cannot not re-use them to evaluate our translation systems after we update their parameters or change their behavior in anyway. The reason for this is that altered systems will produce different translations than the ones that we have judged, so our relative rankings of sentences will no longer be applicable. However, the translations of short phrases are more likely to be repeated than the translations of whole sentences.

Therefore if we collect a large number of yes/no judgments for short phrases, we could build up a database that contains information about what fragmentary translations are acceptable for each sentence in our test corpus. When we change our system and want to evaluate it, we do not need to manually evaluate those segments that match against the database, and could instead have people evaluate only those phrasal translations which are new. Accumulating these judgments over time would give a very reliable idea of what alternative translations were allowable. This would be useful because it could alleviate the problems associated with Bleu failing to recognize allowable variation in translation when multiple reference translations are not available (Callison-Burch et al., 2006). A large database of human judgments might also be useful as an objective function for minimum error rate training (Och, 2003) or in other system development tasks.

## 8 Conclusions

Similar to 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. One important aspect in which this year's

shared task differed from previous years was the introduction of an additional newswire test set that was different in nature from the training data. We also added new language pairs to our evaluation: Hungarian-English and German-Spanish.

As in previous years we were pleased to notice an increase in the number of participants. This year we received submissions from 23 groups from 18 institutions. In addition, we evaluated seven commercial rule-based MT systems.

The goal of this shared-task is two-fold: First we want to compare state-of-the-art machine translation systems, and secondly we aim to measure to what extent different evaluation metrics can be used to assess MT quality.

With respect to MT quality we noticed that the introduction of test sets from a different domain did have an impact on the ranking of systems. We observed that rule-based systems generally did better on the News test set. Overall, it cannot be concluded that one approach clearly outperforms other approaches, as systems performed differently on the various translation tasks. One general observation is that for the tasks where statistical combination approaches participated, they tended to score relatively high, in particular with respect to Bleu.

With respect to measuring the correlation between automated evaluation metrics and human judgments we found that Gimenez and Marquez (2008)'s metrics had very strong system-level correlation when translating into English. When translating from English into French, German, and Spanish methods which utilized part-of-speech resulted in the highest correlations with human judgments. These results suggest that integrating non-lexical information has the potential to strengthen automatic evaluation metrics.

Finally, we investigated inter- and intra-annotator agreement of human judgments using Kappa coefficients. We noticed that ranking whole sentences results in relatively low Kappa coefficients, meaning that there is only fair agreement between the assessors. Constituent ranking and acceptability judgments on the other hand show moderate and substantial inter-annotator agreement, respectively. Intra-annotator agreement was substantial to almost perfect, except for the sentence ranking assessment where agreement was only moderate. Although it



is difficult to draw exact conclusions from this, one might wonder whether the sentence ranking task is simply too complex, involving too many aspects according to which translations can be ranked.

The huge wealth of the data generated by this workshop, including the human judgments, system translations and automatic scores, is available at <http://www.statmt.org/wmt08/> for other researchers to analyze.

## Acknowledgments

This work was supported in parts by the EuroMatrix project funded by the European Commission (6th Framework Programme), the GALE program of the US Defense Advanced Research Projects Agency, Contract No. HR0011-06-C-0022, and the US National Science Foundation under grant IIS-0713448.

We are grateful to Abhaya Agarwal, John Henderson, Rebecca Hwa, Alon Lavie, Mark Przybocki, Stuart Shieber, and David Smith for discussing different possibilities for calculating the sentence-level correlation of automatic evaluation metrics with human judgments in absence of absolute scores. Any errors in design remain the responsibility of the authors.

Thank you to Eckhard Bick for parsing the Spanish test set. See <http://beta.visl.sdu.dk> for more information about the constraint-based parser. Thanks to Greg Hanneman and Antti-Veikko Rosti for applying their system combination algorithms to our data.

## References

- Abhaya Agarwal and Alon Lavie. 2008. Meteor, M-BLEU and M-TER: Evaluation metrics for high-correlation with human rankings of machine translation output. In *Proceedings of the Third Workshop on Statistical Machine Translation*, pages 115–118, Columbus, Ohio, June. Association for Computational Linguistics.
- Joshua Albrecht and Rebecca Hwa. 2007a. A re-examination of machine learning approaches for sentence-level mt evaluation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics (ACL-2007)*, Prague, Czech Republic.
- Joshua Albrecht and Rebecca Hwa. 2007b. Regression for sentence-level mt evaluation with pseudo references. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics (ACL-2007)*, Prague, Czech Republic.
- Joshua Albrecht and Rebecca Hwa. 2008. The role of pseudo references in MT evaluation. In *Proceedings of the Third Workshop on Statistical Machine Translation*, pages 187–190, Columbus, Ohio, June. Association for Computational Linguistics.
- Abhishek Arun and Frank Keller. 2005. Lexicalization in crosslinguistic probabilistic parsing: The case of French. In *Proceedings of ACL*.
- Amittai Axelrod, Mei Yang, Kevin Duh, and Katrin Kirchhoff. 2008. The University of Washington machine translation system for ACL WMT 2008. In *Proceedings of the Third Workshop on Statistical Machine Translation*, pages 123–126, Columbus, Ohio, June. Association for Computational Linguistics.
- Nguyen Bach, Qin Gao, and Stephan Vogel. 2008. Improving word alignment with language model based confidence scores. In *Proceedings of the Third Workshop on Statistical Machine Translation*, pages 151–154, Columbus, Ohio, June. Association for Computational Linguistics.
- Eckhard Bick. 2006. A constraint grammar-based parser for Spanish. In *Proceedings of the 4th Workshop on Information and Human Language Technology (IHLT-2006)*, Ribeiro Preto, Brazil.
- Dan Bikel. 2002. Design of a multi-lingual, parallel-processing statistical parsing engine. In *Proceedings of Second International Conference on Human Language Technology Research (HLT-02)*, San Diego, California.
- Graeme Blackwood, Adrià de Gispert, Jamie Brunning, and William Byrne. 2008. European language translation with weighted finite state transducers: The CUED MT system for the 2008 ACL workshop on SMT. In *Proceedings of the Third Workshop on Statistical Machine Translation*, pages 131–134, Columbus, Ohio, June. Association for Computational Linguistics.
- Ondřej Bojar and Jan Hajič. 2008. Phrase-based and deep syntactic English-to-Czech statistical machine translation. In *Proceedings of the Third Workshop on Statistical Machine Translation*, pages 143–146, Columbus, Ohio, June. Association for Computational Linguistics.
- Chris Callison-Burch, Miles Osborne, and Philipp Koehn. 2006. Re-evaluating the role of Bleu in machine translation research. In *11th Conference of the European Chapter of the Association for Computational Linguistics (EACL-2006)*, Trento, Italy.

- 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*, pages 136–158, Prague, Czech Republic, June. Association for Computational Linguistics.
- Jean Carletta. 1996. Assessing agreement on classification tasks: The kappa statistic. *Computational Linguistics*, 22(2):249–254.
- Daniel Déchelotte, Gilles Adda, Alexandre Allauzen, Hélène Bonneau-Maynard, Olivier Galibert, Jean-Luc Gauvain, Philippe Langlais, and François Yvon. 2008. Limsi’s statistical translation systems for WMT’08. In *Proceedings of the Third Workshop on Statistical Machine Translation*, pages 107–110, Columbus, Ohio, June. Association for Computational Linguistics.
- Amit Dubey. 2005. What to do when lexicalization fails: parsing German with suffix analysis and smoothing. In *Proceedings of ACL*.
- Loïc Dugast, Jean Senellart, and Philipp Koehn. 2008. Can we relearn an RBMT system? In *Proceedings of the Third Workshop on Statistical Machine Translation*, pages 175–178, Columbus, Ohio, June. Association for Computational Linguistics.
- Kevin Duh. 2008. Ranking vs. regression in machine translation evaluation. In *Proceedings of the Third Workshop on Statistical Machine Translation*, pages 191–194, Columbus, Ohio, June. Association for Computational Linguistics.
- Christopher J. Dyer. 2007. The ‘noisier channel’: translation from morphologically complex languages. In *Proceedings of the ACL-2007 Workshop on Statistical Machine Translation (WMT-07)*, Prague, Czech Republic.
- Andreas Eisele, Christian Federmann, Hervé Saint-Amand, Michael Jellinghaus, Teresa Herrmann, and Yu Chen. 2008. Using Moses to integrate multiple rule-based machine translation engines into a hybrid system. In *Proceedings of the Third Workshop on Statistical Machine Translation*, pages 179–182, Columbus, Ohio, June. Association for Computational Linguistics.
- Jesus Gimenez and Lluís Marquez. 2008. A smorgasbord of features for automatic MT evaluation. In *Proceedings of the Third Workshop on Statistical Machine Translation*, pages 195–198, Columbus, Ohio, June. Association for Computational Linguistics.
- Kevin Gimpel and Noah A. Smith. 2008. Rich source-side context for statistical machine translation. In *Proceedings of the Third Workshop on Statistical Machine Translation*, pages 9–17, Columbus, Ohio, June. Association for Computational Linguistics.
- Greg Hanneman, Edmund Huber, Abhaya Agarwal, Vamsi Ambati, Alok Parlikar, Erik Peterson, and Alon Lavie. 2008. Statistical transfer systems for French-English and German-English machine translation. In *Proceedings of the Third Workshop on Statistical Machine Translation*, pages 163–166, Columbus, Ohio, June. Association for Computational Linguistics.
- Shyamsundar Jayaraman and Alon Lavie. 2005. Multi-engine machine translation guided by explicit word matching. In *Proceedings of the 10th Annual Conference of the European Association for Machine Translation*, pages 143–152, Budapest, Hungary, May.
- Maxim Khalilov, Adolfo Hernández H., Marta R. Costajussà, Josep M. Crego, Carlos A. Henríquez Q., Patrik Lambert, José A. R. Fonollosa, José B. Mariño, and Rafael E. Banchs. 2008. The TALP-UPC Ngram-based statistical machine translation system for ACL-WMT 2008. In *Proceedings of the Third Workshop on Statistical Machine Translation*, pages 127–130, Columbus, Ohio, June. Association for Computational Linguistics.
- 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 and Josh Schroeder. 2007. Experiments in domain adaptation for statistical machine translation. In *Proceedings of the ACL-2007 Workshop on Statistical Machine Translation (WMT-07)*, Prague, Czech Republic.
- Philipp Koehn, Franz Josef Och, and Daniel Marcu. 2003. Statistical phrase-based translation. In *Proceedings of the Human Language Technology Conference of the North American chapter of the Association for Computational Linguistics (HLT/NAACL-2003)*, Edmonton, Alberta.
- Philipp Koehn, Nicola Bertoldi, Ondrej Bojar, Chris Callison-Burch, Alexandra Constantin, Brooke Cowan, Chris Dyer, Marcello Federico, Evan Herbst, Hieu Hoang, Christine Moran, Wade Shen, and Richard Zens. 2007. Open source toolkit for statistical machine translation: Factored translation models and confusion network decoding. CLSP Summer Workshop Final Report WS-2006, Johns Hopkins University.
- Philipp Koehn, Abhishek Arun, and Hieu Hoang. 2008. Towards better machine translation quality for the German-English language pairs. In *Proceedings of the Third Workshop on Statistical Machine Translation*, pages 139–142, Columbus, Ohio, June. Association for Computational Linguistics.

- Alex Kulesza and Stuart M. Shieber. 2004. A learning approach to improving sentence-level MT evaluation. In *Proceedings of the 10th International Conference on Theoretical and Methodological Issues in Machine Translation*, Baltimore, MD, October 4–6.
- J. R. Landis and G. G. Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics*, 33:159–174.
- LDC. 2005. Linguistic data annotation specification: Assessment of fluency and adequacy in translations. Revision 1.5.
- Preslav Nakov. 2008. Improving English-Spanish statistical machine translation: Experiments in domain adaptation, sentence paraphrasing, tokenization, and recasing. In *Proceedings of the Third Workshop on Statistical Machine Translation*, pages 147–150, Columbus, Ohio, June. Association for Computational Linguistics.
- Vassilina Nikoulina and Marc Dymetman. 2008. Using syntactic coupling features for discriminating phrase-based translations (wmt-08 shared translation task). In *Proceedings of the Third Workshop on Statistical Machine Translation*, pages 159–162, Columbus, Ohio, June. Association for Computational Linguistics.
- Attila Novák, László Tihanyi, and Gábor Prószték. 2008. The MetaMorpho translation system. In *Proceedings of the Third Workshop on Statistical Machine Translation*, pages 111–114, Columbus, Ohio, June. Association for Computational Linguistics.
- Franz Josef Och and Hermann Ney. 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51, March.
- 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.
- Maja Popovic and Hermann Ney. 2007. Word error rates: Decomposition over POS classes and applications for error analysis. In *Proceedings of ACL Workshop on Machine Translation*, Prague, Czech Republic.
- Mark Przybocki and Kay Peterson, editors. 2008. *Proceedings of the 2008 NIST Open Machine Translation Evaluation Workshop*. Arlington, Virginia, March.
- Antti-Veikko Rosti, Spyros Matsoukas, and Richard Schwartz. 2007. Improved word-level system combination for machine translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics (ACL-2007)*, Prague, Czech Republic.
- Antti-Veikko Rosti, Bing Zhang, Spyros Matsoukas, and Richard Schwartz. 2008. Incremental hypothesis alignment for building confusion networks with application to machine translation system combination. In *Proceedings of the Third Workshop on Statistical Machine Translation*, pages 183–186, Columbus, Ohio, June. Association for Computational Linguistics.
- Holger Schwenk, Jean-Baptiste Fouet, and Jean Senellart. 2008. First steps towards a general purpose French/English statistical machine translation system. In *Proceedings of the Third Workshop on Statistical Machine Translation*, pages 119–122, Columbus, Ohio, June. Association for Computational Linguistics.
- Sara Stymne, Maria Holmqvist, and Lars Ahrenberg. 2008. Effects of morphological analysis in translation between German and English. In *Proceedings of the Third Workshop on Statistical Machine Translation*, pages 135–138, Columbus, Ohio, June. Association for Computational Linguistics.
- John Tinsley, Yanjun Ma, Sylwia Ozdowska, and Andy Way. 2008. MaTrEx: The DCU MT system for WMT 2008. In *Proceedings of the Third Workshop on Statistical Machine Translation*, pages 171–174, Columbus, Ohio, June. Association for Computational Linguistics.
- Zhuoran Wang and John Shawe-Taylor. 2008. Kernel regression framework for machine translation: UCL system description for WMT 2008 shared translation task. In *Proceedings of the Third Workshop on Statistical Machine Translation*, pages 155–158, Columbus, Ohio, June. Association for Computational Linguistics.
- Zdenek Zabokrtsky, Jan Ptacek, and Petr Pajas. 2008. TectoMT: Highly modular MT system with tectogramatics used as transfer layer. In *Proceedings of the Third Workshop on Statistical Machine Translation*, pages 167–170, Columbus, Ohio, June. Association for Computational Linguistics.

## A Automatic scores for each system

	BLEU	MBLEU	METEOR-BASELINE	METEOR-RANKING	MTER	SVM-RANK
English-Czech News Commentary Task						
CU-BOJAR	<b>0.15</b>	<b>0.21</b>	<b>0.43</b>	<b>0.35</b>	<b>0.28</b>	<b>4.57</b>
CU-BOJAR-CONTRAST-1	0.04	0.11	0.32	0.25	0.18	0.90
CU-BOJAR-CONTRAST-2	0.14	0.2	0.42	0.34	0.27	2.86
CU-TECTOMT	0.09	0.15	0.37	0.29	0.23	2.13
PC-TRANSLATOR	0.08	0.14	0.35	0.28	0.19	2.09
UEDIN	0.12	0.18	0.4	0.32	0.25	2.28
English-Czech News Task						
CU-BOJAR	<b>0.11</b>	<b>0.18</b>	<b>0.37</b>	<b>0.3</b>	<b>0.18</b>	<b>4.72</b>
CU-BOJAR-CONTRAST-1	0.02	0.10	0.26	0.2	0.12	0.80
CU-BOJAR-CONTRAST-2	0.09	0.16	0.35	0.28	0.15	2.65
CU-TECTOMT	0.06	0.13	0.32	0.25	0.16	2.14
PC-TRANSLATOR	0.08	0.14	0.33	0.26	0.14	2.40
UEDIN	0.08	0.15	0.34	0.27	0.15	2.13

Table 14: Automatic evaluation metric for translations into Czech

	BLEU	MBLEU	METEOR-B	METEOR-R	MTER	POSF4G-AM	POSF4G-GM	POSBLEU	SVM-RANK
English-French News Task									
LIMS1	0.2	0.26	0.16	0.34	0.33	0.48	0.44	0.43	<b>9.74</b>
LIUM-SYSTRAN	<b>0.20</b>	<b>0.26</b>	<b>0.16</b>	0.35	<b>0.34</b>	<b>0.49</b>	<b>0.44</b>	<b>0.44</b>	7.38
LIUM-SYSTRAN-CONTRAST	0.20	0.26	0.16	<b>0.35</b>	0.34	0.48	0.44	0.44	7.02
RBMT1	0.13	0.19	0.12	0.28	0.24	0.42	0.37	0.35	5.46
RBMT3	0.17	0.23	0.14	0.31	0.31	0.45	0.4	0.40	5.60
RBMT4	0.19	0.24	0.15	0.33	0.32	0.48	0.43	0.43	6.80
RBMT5	0.17	0.23	0.14	0.32	0.31	0.47	0.42	0.42	6.15
RBMT6	0.16	0.22	0.13	0.32	0.3	0.46	0.40	0.41	5.60
SAAR	0.15	0.22	0.15	0.33	0.28	0.46	0.41	0.42	6.12
SAAR-CONTRAST	0.17	0.23	0.15	0.33	0.30	0.47	0.42	0.41	5.50
UEDIN	0.16	0.23	0.14	0.32	0.32	0.44	0.39	0.38	4.79
XEROX	0.13	0.2	0.12	0.29	0.29	0.41	0.34	0.34	3.91
XEROX-CONTRAST	0.13	0.2	0.12	0.29	0.29	0.41	0.35	0.35	3.86
English-French Europarl Task									
LIMS1	0.32	0.36	0.24	0.42	0.44	0.56	0.53	0.53	<b>8.84</b>
LIUM-SYSTRAN	<b>0.32</b>	<b>0.36</b>	<b>0.24</b>	<b>0.42</b>	<b>0.45</b>	<b>0.56</b>	<b>0.53</b>	<b>0.53</b>	7.46
LIUM-SYSTRAN-CONTRAST	0.31	0.36	0.23	0.42	0.44	0.56	0.52	0.53	6.69
RBMT1	0.15	0.20	0.13	0.29	0.26	0.44	0.4	0.37	3.89
RBMT3	0.18	0.24	0.15	0.34	0.33	0.47	0.42	0.43	4.13
RBMT4	0.2	0.25	0.17	0.35	0.35	0.5	0.45	0.45	4.70
RBMT5	0.12	0.16	0.09	0.22	0.06	0.37	0.32	0.32	3.01
RBMT6	0.17	0.23	0.14	0.33	0.32	0.47	0.42	0.42	3.93
SAAR	0.26	0.29	0.21	0.41	0.34	0.53	0.49	0.48	7.75
SAAR-CONTRAST	0.28	0.32	0.23	0.41	0.39	0.55	0.51	0.52	6.45
UCL	0.24	0.28	0.19	0.37	0.41	0.49	0.44	0.42	4.16
UEDIN	0.30	0.35	0.23	0.42	0.43	0.54	0.51	0.51	6.56

Table 15: Automatic evaluation metric for translations into French

	BLEU	MBLEU	METEOR-BASELINE	METEOR-RANKING	MTER	POSF4GRAM-AM	POSF4GRAM-GM	POSBLEU	SVM-RANK
English-German News Task									
LIMSI	0.11	0.18	0.19	0.45	0.22	0.36	0.29	0.28	<b>7.83</b>
LIU	0.10	0.17	0.18	0.44	0.24	0.36	0.28	0.27	4.03
RBMT1	0.12	0.18	0.18	0.44	0.22	0.39	0.33	0.32	5.42
RBMT2	0.13	0.19	<b>0.20</b>	<b>0.46</b>	0.24	0.4	0.33	0.33	5.76
RBMT3	0.12	0.18	0.19	0.44	0.24	0.39	0.32	0.32	4.70
RBMT4	<b>0.14</b>	<b>0.19</b>	0.2	0.46	0.25	<b>0.41</b>	<b>0.35</b>	<b>0.34</b>	5.58
RBMT5	0.11	0.17	0.17	0.43	0.21	0.38	0.31	0.31	4.49
RBMT6	0.10	0.16	0.17	0.43	0.2	0.37	0.3	0.29	4.81
SAAR	0.13	0.19	0.19	0.44	<b>0.27</b>	0.38	0.31	0.3	4.04
SAAR-CONTRAST	0.12	0.18	0.18	0.43	0.26	0.37	0.3	0.28	3.71
UEDIN	0.12	0.17	0.18	0.45	0.23	0.37	0.30	0.29	4.37
English-German Europarl Task									
CMU-GIMPEL	0.20	0.24	0.27	<b>0.54</b>	0.32	0.43	0.37	0.37	<b>9.54</b>
LIMSI	0.20	0.24	0.27	0.53	0.32	0.43	0.37	0.37	6.97
LIU	0.2	0.24	0.27	0.53	<b>0.32</b>	0.43	0.38	0.37	6.95
RBMT1	0.11	0.16	0.16	0.42	0.19	0.38	0.32	0.32	5.01
RBMT2	0.12	0.17	0.19	0.46	0.21	0.39	0.32	0.31	5.93
RBMT3	0.11	0.16	0.17	0.43	0.21	0.38	0.31	0.30	4.75
RBMT4	0.12	0.17	0.18	0.45	0.22	0.41	0.34	0.33	5.42
RBMT5	0.1	0.14	0.16	0.42	0.19	0.39	0.32	0.31	4.42
RBMT6	0.09	0.14	0.15	0.42	0.18	0.38	0.30	0.29	4.40
SAAR	0.20	0.25	0.26	0.53	0.32	0.43	0.38	0.37	6.67
SAAR-CONTRAST	0.2	0.24	0.26	0.52	0.31	0.43	0.37	0.37	6.35
UCL	0.16	0.20	0.23	0.49	0.31	0.4	0.33	0.31	5.12
UEDIN	<b>0.21</b>	<b>0.25</b>	<b>0.27</b>	0.54	0.32	<b>0.44</b>	<b>0.38</b>	<b>0.38</b>	7.02
English-Spanish News Task									
CMU-SMT	0.19	0.24	0.25	0.34	0.32	0.32	0.25	0.26	<b>8.34</b>
LIMSI	0.19	0.25	0.26	0.34	0.34	0.33	0.26	0.26	5.92
RBMT1	0.16	0.22	0.23	0.32	0.30	0.31	0.23	0.23	5.36
RBMT3	0.19	0.24	0.25	0.33	0.34	0.33	0.26	0.26	5.42
RBMT4	<b>0.21</b>	0.26	<b>0.26</b>	0.34	0.35	0.34	0.28	<b>0.28</b>	6.36
RBMT5	0.18	0.24	0.25	0.33	0.32	0.33	0.26	0.26	5.84
RBMT6	0.19	0.24	0.24	0.33	0.33	0.32	0.25	0.26	5.42
SAAR	0.20	<b>0.27</b>	0.26	0.34	<b>0.37</b>	<b>0.34</b>	<b>0.28</b>	0.28	5.04
SAAR-CONTRAST	0.2	0.26	0.25	0.34	0.37	0.34	0.27	0.27	4.86
UCB	0.20	0.26	0.26	<b>0.34</b>	0.34	0.33	0.26	0.27	5.70
UEDIN	0.18	0.25	0.25	0.33	0.35	0.33	0.26	0.26	4.30
UPC	0.18	0.23	0.24	0.32	0.35	0.32	0.25	0.24	3.97
English-Spanish Europarl Task									
CMU-SMT	<b>0.32</b>	0.36	0.33	0.42	0.45	0.40	0.35	0.36	<b>0.10</b>
LIMSI	0.31	0.36	0.33	0.42	0.45	0.4	0.35	0.35	7.80
RBMT1	0.16	0.22	0.24	0.32	0.31	0.32	0.25	0.25	4.47
RBMT3	0.20	0.25	0.25	0.34	0.35	0.33	0.27	0.27	4.66
RBMT4	0.21	0.25	0.26	0.34	0.36	0.34	0.28	0.28	4.85
RBMT5	0.18	0.24	0.25	0.34	0.33	0.34	0.27	0.27	5.03
RBMT6	0.18	0.23	0.25	0.33	0.33	0.33	0.26	0.26	4.57
SAAR	0.31	0.35	0.33	0.41	0.44	0.40	0.35	0.35	7.59
SAAR-CONTRAST	0.30	0.34	0.33	0.41	0.44	0.4	0.34	0.35	7.42
UCL	0.25	0.29	0.29	0.37	0.43	0.36	0.29	0.29	4.67
UEDIN	0.32	<b>0.36</b>	0.33	0.42	0.45	0.40	0.35	0.35	7.25
UPC	0.30	0.34	0.32	0.40	<b>0.46</b>	0.4	0.35	0.34	6.18
UW	0.32	0.36	<b>0.33</b>	<b>0.42</b>	0.45	0.40	0.35	0.35	7.36
UW-CONTRAST	0.32	0.35	0.33	0.42	0.45	<b>0.40</b>	<b>0.35</b>	<b>0.36</b>	7.21

Table 16: Automatic evaluation metric for translations into German and Spanish

	DP	DR	SR	ULC	ULCH	BLEU	MBLEU	METEOR-BASELINE	METEOR-RANKING	MTER	POSF4GRAM-AM	POSF4GRAM-GM	POSBLEU	SVM-RANK
Spanish-English Europarl Task														
CMU-SMT	0.34	<b>0.44</b>	0.26	0.29	0.33	<b>0.33</b>	0.39	0.59	0.48	0.51	0.52	0.48	0.48	<b>9.72</b>
CUED	0.33	0.43	0.25	0.29	0.33	0.32	0.38	0.59	0.48	0.50	0.51	0.47	0.47	7.41
CUED-CONTRAST	0.34	0.44	0.26	0.29	0.33	0.33	0.39	0.59	0.48	0.51	0.52	0.48	0.48	7.00
DCU	0.34	0.43	0.25	0.29	0.33	0.32	0.38	0.59	0.48	0.50	0.51	0.47	0.48	6.78
LIMSI	<b>0.34</b>	0.44	0.26	<b>0.29</b>	<b>0.33</b>	0.33	<b>0.39</b>	<b>0.59</b>	<b>0.48</b>	<b>0.51</b>	0.52	0.48	<b>0.48</b>	6.73
RBMT3	0.26	0.37	0.19	0.22	0.27	0.19	0.26	0.51	0.41	0.36	0.45	0.4	0.39	5.46
RBMT4	0.26	0.37	0.19	0.22	0.27	0.18	0.26	0.52	0.42	0.36	0.45	0.39	0.38	5.57
RBMT5	0.25	0.36	0.18	0.22	0.27	0.18	0.25	0.51	0.41	0.36	0.44	0.39	0.38	4.74
RBMT6	0.24	0.34	0.18	0.21	0.26	0.17	0.25	0.51	0.41	0.36	0.44	0.38	0.37	4.71
SAAR	0.34	0.44	<b>0.26</b>	0.29	0.33	0.32	0.39	0.59	0.48	0.51	<b>0.52</b>	<b>0.49</b>	0.48	6.30
SAAR-CONTRAST	0.33	0.43	0.25	0.28	0.33	0.30	0.37	0.59	0.48	0.47	0.51	0.47	0.46	7.33
UCL	0.29	0.4	0.21	0.25	0.29	0.25	0.32	0.55	0.43	0.47	0.47	0.42	0.4	4.02
UEDIN	0.34	0.44	0.26	0.29	0.33	0.33	0.39	0.59	0.48	0.50	0.52	0.48	0.48	6.61
UPC	0.33	0.43	0.25	0.28	0.33	0.32	0.38	0.59	0.48	0.5	0.52	0.48	0.48	6.82
French-English News Task														
BBN-COMBO	<b>0.27</b>	0.37	<b>0.2</b>	<b>0.23</b>	<b>0.28</b>	0.21	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a
CMU-COMBO	0.26	0.36	0.18	0.22	0.27	0.19	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a
CMU-COMBO-CONTRAST	n/a	n/a	n/a	n/a	n/a	0.19	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a
CMU-STATXFER	0.21	0.32	0.14	0.19	0.23	0.14	0.22	0.48	0.39	0.28	0.38	0.32	0.30	<b>9.91</b>
CMU-STATXFER-CONTRAST	0.21	0.30	0.14	0.18	0.23	0.14	0.21	0.47	0.38	0.26	0.38	0.31	0.29	6.47
CUED	0.25	0.35	0.17	0.21	0.26	0.18	0.27	0.51	0.41	0.37	0.41	0.35	0.34	6.34
CUED-CONTRAST	0.26	0.37	0.18	0.22	0.27	0.19	0.28	<b>0.52</b>	<b>0.42</b>	0.38	0.42	0.37	0.36	6.29
LIMSI	0.26	0.37	0.18	0.22	0.27	0.20	0.28	0.51	0.40	0.40	0.43	0.38	0.37	5.75
LIUM-SYSTRAN	0.27	<b>0.38</b>	0.19	0.23	0.27	0.21	0.29	0.51	0.41	0.41	<b>0.44</b>	<b>0.39</b>	<b>0.38</b>	6.32
LIUM-SYSTRAN-CONTRAST	0.27	0.38	0.19	0.23	0.28	<b>0.21</b>	<b>0.29</b>	0.51	0.41	<b>0.41</b>	0.44	0.39	0.38	5.93
RBMT3	0.24	0.36	0.17	0.21	0.26	0.16	0.24	0.49	0.40	0.29	0.42	0.36	0.34	7.61
RBMT4	0.25	0.37	0.17	0.21	0.26	0.17	0.25	0.49	0.4	0.33	0.42	0.36	0.35	6.17
RBMT5	0.25	0.37	0.18	0.22	0.27	0.18	0.25	0.51	0.41	0.33	0.43	0.37	0.36	6.97
RBMT6	0.24	0.36	0.17	0.21	0.26	0.16	0.24	0.49	0.39	0.30	0.41	0.35	0.34	6.51
SAAR	0.24	0.14	0.17	0.19	0.22	0.15	0.24	0.47	0.37	0.39	0.39	0.32	0.31	3.22
SAAR-CONTRAST	0.26	0.36	0.18	0.22	0.27	0.17	0.27	0.51	0.41	0.36	0.41	0.35	0.35	6.01
UEDIN	0.25	0.36	0.17	0.21	0.26	0.18	0.26	0.51	0.41	0.35	0.42	0.36	0.35	5.97
UEDIN-COMBO	0.26	0.36	0.18	0.23	0.27	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a
French-English Europarl Task														
CMU-STATXFER	0.24	0.34	0.18	0.22	0.26	0.2	0.26	0.52	0.42	0.37	0.42	0.36	0.35	9.85
CMU-STATXFER-CONTRAST	0.25	0.34	0.19	0.22	0.26	0.2	0.26	0.53	0.42	0.38	0.42	0.36	0.35	7.10
CUED	0.34	0.44	0.26	0.29	0.33	0.32	0.38	0.59	0.48	0.50	0.51	0.47	0.47	<b>0.11</b>
CUED-CONTRAST	0.34	0.44	0.26	0.29	0.34	0.32	0.39	0.59	0.48	0.51	0.51	0.47	0.47	9.34
DCU	0.33	0.43	0.25	0.28	0.33	0.31	0.37	0.58	0.47	0.49	0.50	0.46	0.46	9.16
LIMSI	0.34	0.44	0.26	0.29	0.34	0.33	0.39	0.59	0.48	0.51	0.52	0.48	0.48	9.59
LIUM-SYSTRAN	<b>0.35</b>	<b>0.45</b>	<b>0.27</b>	<b>0.3</b>	<b>0.34</b>	<b>0.33</b>	<b>0.39</b>	<b>0.59</b>	<b>0.48</b>	<b>0.51</b>	<b>0.52</b>	<b>0.48</b>	<b>0.49</b>	9.75
LIUM-SYSTRAN-CONTRAST	0.34	0.44	0.26	0.29	0.34	0.33	0.39	0.59	0.48	0.50	0.52	0.48	0.48	9.23
RBMT3	0.25	0.36	0.10	0.20	0.24	0.17	0.25	0.51	0.41	0.35	0.43	0.37	0.36	7.36
RBMT4	0.27	0.36	0.19	0.22	0.27	0.18	0.26	0.51	0.41	0.37	0.43	0.38	0.37	5.92
RBMT5	0.27	0.38	0.21	0.23	0.28	0.20	0.28	0.53	0.43	0.4	0.45	0.4	0.39	7.20
RBMT6	0.24	0.35	0.18	0.21	0.26	0.16	0.24	0.5	0.40	0.35	0.42	0.36	0.35	5.96
SAAR	0.32	0.41	0.23	0.27	0.31	0.27	0.33	0.54	0.43	0.49	0.49	0.44	0.41	4.76
SAAR-CONTRAST	0.33	0.43	0.25	0.28	0.33	0.3	0.36	0.58	0.48	0.47	0.51	0.47	0.46	0.10
SYSTRAN	0.3	0.4	0.23	0.26	0.30	0.26	0.34	0.55	0.45	0.46	0.48	0.43	0.43	7.01
UCL	0.3	0.40	0.22	0.26	0.3	0.26	0.32	0.55	0.44	0.47	0.47	0.42	0.41	6.35
UEDIN	0.34	0.44	0.26	0.29	0.33	0.33	0.39	0.59	0.48	0.50	0.52	0.48	0.48	9.41

Table 17: Automatic evaluation metric for translations into English

	DP	DR	SR	ULC	ULCH	BLEU	MBLEU	METEOR-BASELINE	METEOR-RANKING	MTER	POSF4GRAM-AM	POSF4GRAM-GM	POSBLEU	SVM-RANK
Czech-English News Commentary Task														
DCU	0.25	<b>0.34</b>	0.18	0.22	0.27	0.21	0.29	0.54	0.44	0.42	0.42	0.36	0.36	<b>2.45</b>
SYSTRAN	0.19	0.28	0.12	0.17	0.21	0.15	0.23	0.45	0.36	0.34	0.36	0.29	0.29	0.76
UEDIN	0.24	0.31	0.16	0.21	0.25	0.22	0.30	0.54	0.44	0.43	0.41	0.35	0.35	1.37
UMD	<b>0.26</b>	<b>0.34</b>	<b>0.19</b>	<b>0.23</b>	<b>0.28</b>	<b>0.24</b>	<b>0.33</b>	<b>0.56</b>	<b>0.45</b>	<b>0.49</b>	<b>0.44</b>	<b>0.39</b>	<b>0.38</b>	1.41
Czech-English News Task														
DCU	0.19	<b>0.30</b>	<b>0.13</b>	0.17	<b>0.22</b>	0.12	<b>0.22</b>	<b>0.45</b>	<b>0.35</b>	0.32	0.36	0.28	<b>0.28</b>	<b>1.78</b>
UEDIN	0.19	0.28	0.12	0.17	0.21	0.12	0.21	0.44	0.34	0.32	0.35	0.27	0.27	0.65
UMD	<b>0.2</b>	0.29	0.12	<b>0.18</b>	0.22	<b>0.13</b>	0.22	0.44	0.34	<b>0.36</b>	<b>0.36</b>	<b>0.29</b>	0.27	0.52
German-English News Task														
BBN-COMBO	0.23	0.34	0.14	0.21	0.25	<b>0.18</b>	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a
CMU-STATXFER	0.16	0.27	0.09	0.15	0.19	0.11	0.18	0.43	0.34	0.25	0.33	0.25	0.24	<b>7.84</b>
LIMSI	0.22	0.33	0.13	0.19	0.23	0.17	0.25	0.47	0.37	<b>0.36</b>	0.4	0.33	0.32	5.58
LIU	0.21	0.32	0.06	0.18	0.22	0.15	0.24	0.48	0.38	0.33	0.38	0.31	0.31	5.51
RBMT1	0.22	0.33	0.14	0.19	0.23	0.14	0.22	0.44	0.35	0.28	0.37	0.31	0.30	6.13
RBMT2	0.24	0.37	0.17	0.21	0.26	0.15	0.24	0.5	0.40	0.31	0.4	0.33	0.32	7.14
RBMT3	0.24	0.37	0.16	0.21	0.26	0.16	0.24	0.49	0.4	0.32	0.41	0.34	0.34	6.97
RBMT4	<b>0.25</b>	<b>0.38</b>	<b>0.17</b>	<b>0.21</b>	<b>0.27</b>	0.16	0.25	<b>0.50</b>	<b>0.40</b>	0.34	<b>0.41</b>	<b>0.35</b>	<b>0.34</b>	7.03
RBMT5	0.23	0.36	0.15	0.20	0.25	0.15	0.23	0.48	0.39	0.32	0.4	0.33	0.32	5.94
RBMT6	0.22	0.34	0.14	0.19	0.24	0.14	0.22	0.47	0.38	0.31	0.39	0.32	0.31	5.65
SAAR	0.22	0.33	0.14	0.2	0.24	0.15	0.24	0.47	0.37	0.36	0.39	0.32	0.31	4.67
SAAR-CONTRAST	0.24	0.35	0.16	0.21	0.25	0.17	<b>0.26</b>	0.5	0.4	0.36	0.4	0.33	0.33	5.80
SAAR-CONTRAST-2	0.21	0.33	0.14	0.19	0.23	0.15	0.24	0.47	0.37	0.36	0.39	0.32	0.31	4.80
UEDIN	0.23	0.34	0.09	0.19	0.23	0.16	0.25	0.48	0.39	0.35	0.4	0.33	0.33	5.72
German-English Europarl Task														
CMU-STATXFER	0.2	0.31	0.12	0.19	0.22	0.17	0.23	0.49	0.39	0.34	0.39	0.32	0.31	7.11
LIMSI	0.28	0.38	0.18	0.24	0.28	0.27	0.33	0.55	0.44	0.43	0.47	0.42	0.42	<b>8.04</b>
LIU	0.28	0.39	0.09	0.23	0.26	0.27	0.33	0.55	0.44	0.44	0.47	0.43	0.43	7.46
RBMT1	0.21	0.3	0.14	0.18	0.22	0.12	0.19	0.42	0.33	0.27	0.36	0.30	0.28	4.61
RBMT2	0.24	0.35	0.16	0.20	0.25	0.14	0.23	0.49	0.39	0.32	0.39	0.33	0.32	5.42
RBMT3	0.24	0.35	0.16	0.20	0.25	0.15	0.23	0.48	0.39	0.32	0.40	0.34	0.33	5.43
RBMT4	0.24	0.36	0.15	0.20	0.25	0.14	0.23	0.49	0.39	0.34	0.41	0.34	0.34	5.11
RBMT5	0.23	0.34	0.15	0.2	0.24	0.14	0.22	0.48	0.38	0.33	0.4	0.33	0.32	4.55
RBMT6	0.22	0.33	0.13	0.18	0.23	0.13	0.21	0.47	0.37	0.31	0.38	0.31	0.31	4.08
SAAR	0.29	0.39	0.19	0.25	0.28	0.27	0.33	0.55	0.44	0.43	0.47	0.42	0.42	7.32
SAAR-CONTRAST	0.28	0.37	0.18	0.24	0.28	0.26	0.32	0.54	0.43	0.43	0.47	0.42	0.42	6.77
UCL	0.24	0.36	0.16	0.22	0.25	0.2	0.25	0.49	0.39	0.41	0.42	0.35	0.32	4.26
UEDIN	<b>0.30</b>	<b>0.41</b>	<b>0.20</b>	<b>0.26</b>	<b>0.3</b>	<b>0.28</b>	<b>0.34</b>	<b>0.56</b>	<b>0.45</b>	<b>0.45</b>	<b>0.48</b>	<b>0.44</b>	<b>0.44</b>	7.96
Spanish-English News Task														
CMU-SMT	0.24	0.35	0.17	0.21	0.25	0.18	0.26	0.48	0.38	0.39	0.41	0.35	0.34	<b>8.00</b>
CUED	0.25	0.36	0.17	0.21	0.26	0.19	0.28	0.50	0.40	0.38	0.42	0.36	0.36	6.03
CUED-CONTRAST	0.26	0.37	0.18	0.22	0.27	<b>0.21</b>	<b>0.3</b>	<b>0.52</b>	0.42	0.39	0.44	0.38	<b>0.38</b>	6.27
LIMSI	0.26	0.37	0.18	0.22	0.27	0.20	0.28	0.50	0.4	<b>0.41</b>	0.43	0.38	0.37	4.93
RBMT3	0.25	0.38	0.17	0.22	0.27	0.18	0.26	0.50	0.41	0.32	0.43	0.38	0.36	7.54
RBMT4	0.26	<b>0.38</b>	0.18	0.22	0.27	0.18	0.26	0.51	<b>0.42</b>	0.32	<b>0.44</b>	<b>0.39</b>	0.37	7.81
RBMT5	0.26	0.38	0.08	0.20	0.25	0.2	0.27	0.51	0.42	0.33	0.44	0.38	0.37	6.89
RBMT6	0.25	0.36	0.17	0.21	0.26	0.18	0.25	0.51	0.41	0.33	0.43	0.37	0.36	6.83
SAAR	0.26	0.37	0.19	0.22	0.27	0.19	0.29	0.51	0.41	0.39	0.43	0.37	0.37	5.23
SAAR-CONTRAST	0.26	0.37	0.18	0.22	0.27	0.19	0.28	0.51	0.41	0.37	0.42	0.37	0.36	5.95
UCB	0.25	0.35	0.17	0.21	0.26	0.19	0.27	0.5	0.39	0.39	0.42	0.36	0.35	4.40
UEDIN	0.24	0.35	0.17	0.21	0.26	0.18	0.27	0.50	0.40	0.36	0.41	0.35	0.34	5.07
UEDIN-COMBO	<b>0.27</b>	0.36	<b>0.19</b>	<b>0.23</b>	<b>0.27</b>	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a
UPC	0.25	0.36	0.17	0.21	0.26	0.19	0.26	0.49	0.39	0.4	0.43	0.37	0.36	4.38

Table 18: Automatic evaluation metric for translations into English

## B Break down of correlation for each task

	RANK	CONST	YES/NO	DP	DR	SR	ULC	ULCH	BLEU	MBLEU	METEOR-BASELINE	METEOR-RANKING	MTER	POSF4GRAM-AM	POSF4GRAM-GM	POSBLEU	SVM-RANK
All-English News Task																	
RANK	1	n/a	n/a	.83	.73	.83	.83	<b>.87</b>	.71	.7	.82	.79	.41	.79	.8	.80	.25
French-English News Task																	
RANK	1	.69	.63	<b>.92</b>	.83	.89	.90	.90	.81	.80	.88	.80	.57	.87	.9	.9	-.21
CONST	—	1	.81	.83	.52	.81	.86	.81	<b>.93</b>	.9	.76	.64	.73	.69	.72	.85	-.52
YES/NO	—	—	1	.71	.57	.76	.77	.74	<b>.79</b>	.75	.67	.59	.62	.66	.67	.79	-.26
French-English Europarl Task																	
RANK	1	.95	.9	.94	.95	.93	<b>.95</b>	.93	.92	.90	.88	.87	.92	.94	.94	.91	.50
CONST	—	1	.91	.97	.97	<b>.98</b>	.98	.97	.97	.96	.97	.95	.96	.97	.97	.96	.56
YES/NO	—	—	1	.94	.94	.94	.96	.96	<b>.97</b>	.92	.93	.92	.95	.95	.97	.97	.47
German-English News Task																	
RANK	1	.56	.56	.85	.93	.92	.85	<b>.95</b>	.12	.09	.83	.89	-.11	.63	.60	.58	.36
CONST	—	1	.48	.54	.48	.59	<b>.66</b>	.57	.64	.65	.61	.52	.51	.57	.63	.56	-.02
YES/NO	—	—	1	.68	.61	.69	.73	.67	.60	.41	.54	.56	.33	.79	<b>.83</b>	.70	.08
German-English Europarl Task																	
RANK	1	.63	.81	<b>.76</b>	.59	.46	.57	.60	.30	.39	.40	.66	.25	.53	.53	.64	.35
CONST	—	1	.78	.87	<b>.92</b>	.51	.83	.86	.69	.69	.76	.80	.69	.88	.88	.88	.61
YES/NO	—	—	1	<b>.88</b>	.77	.48	.77	.78	.66	.67	.64	.86	.58	.74	.74	.85	.78
Spanish-English News Task																	
RANK	1	-.07	.44	.75	.76	.68	.71	<b>.81</b>	.19	.01	.66	.63	-.12	.73	.76	.66	.36
CONST	—	1	.66	-.03	-.44	.29	.29	.14	.45	.66	-.11	-.33	<b>.77</b>	-.37	-.34	.16	-.58
YES/NO	—	—	1	.29	.05	<b>.73</b>	.64	.52	.48	.47	.09	-.11	.71	.06	.1	.39	-.43
Spanish-English Europarl Task																	
RANK	1	.69	.76	.78	.73	.73	.8	.77	.78	.79	.83	.84	.77	.73	.73	.80	<b>.87</b>
CONST	—	1	.68	.76	.77	.75	.69	.73	.64	.67	.64	.68	.73	<b>.78</b>	<b>.78</b>	.73	.56
YES/NO	—	—	1	.94	.93	.95	.96	.95	<b>.98</b>	.97	.91	.91	.95	.94	.94	<b>.98</b>	.69

Table 19: Correlation of automatic evaluation metrics with the three types of human judgments for translation into English



	RANK	CONST	YES/NO	BLEU	MBLEU	METEOR-BASELINE	METEOR-RANKING	MTER	POSF4GRAM-AM	POSF4GRAM-GM	POSBLEU	SVM-RANK
English-French News Task												
RANK	1	.52	.48	<b>.73</b>	.62	.3	.47	.56	.69	.69	.66	.72
CONST	—	1	.35	.49	.47	.39	.49	.24	<b>.59</b>	<b>.59</b>	.58	.45
YES/NO	—	—	1	.81	<b>.92</b>	.71	.73	.78	.73	.73	.76	.76
English-French Europarl Task												
RANK	1	.98	.88	.95	.95	.95	.95	.90	<b>.97</b>	<b>.97</b>	.93	.93
CONST	—	1	.94	.98	.98	.98	.98	.93	<b>1</b>	<b>1</b>	.97	.91
YES/NO	—	—	1	<b>.97</b>	<b>.97</b>	<b>.97</b>	<b>.97</b>	.92	.95	.95	.92	.83
English-German News Task												
RANK	1	.57	.71	.58	.42	.43	.13	.25	<b>.90</b>	<b>.90</b>	<b>.90</b>	.32
CONST	—	1	.78	.75	<b>.83</b>	.82	.52	.60	.72	.72	.72	.58
YES/NO	—	—	1	.62	.54	.51	.36	.23	.75	.75	.75	<b>.76</b>
English-German Europarl Task												
RANK	1	.28	.57	.36	.36	.42	.39	.26	.38	.38	.50	<b>.56</b>
CONST	—	1	.87	.88	.88	.91	.90	<b>.93</b>	.88	.88	.80	.85
YES/NO	—	—	1	.89	.89	.96	.96	.84	.86	.86	.87	<b>.98</b>
English-Spanish News Task												
RANK	1	-.30	.49	-.04	-.47	-.25	-.29	-.33	-.19	-.19	-.07	<b>.02</b>
CONST	—	1	.43	<b>.79</b>	.61	.64	.56	.2	.59	.59	.52	.56
YES/NO	—	—	1	.52	.41	.43	.31	.13	.65	.65	<b>.72</b>	.16
English-Spanish Europarl Task												
RANK	1	.90	.63	.8	.83	<b>.84</b>	.83	.73	.79	.79	.76	.80
CONST	—	1	.73	.84	<b>.86</b>	.81	.8	.74	.84	.83	.84	<b>.86</b>
YES/NO	—	—	1	.68	.75	.66	.67	<b>.90</b>	.67	.66	.73	.68

Table 20: Correlation of automatic evaluation metrics with the three types of human judgments for translation into other languages

## C Pairwise system comparisons by human judges

The following tables show pairwise comparisons between systems for each language pair, test set, and manual evaluation type. The numbers in each of the tables' cells indicate the percent of 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 complimentary 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  $p \leq 0.05$  and  $\dagger$  indicates statistical significance at  $p \leq 0.01$ , according to the Sign Test.

	BBN-CMB	CMU-CMB	CMU-XFR	CUED	CUED-C	LIMSI	LIUM-SYS	LIUM-SYS-C	RBMT3	RBMT4	RBMT5	RBMT6	SAAR	SAAR-C	UEDIN	UEDIN-CMB
BBN-CMB		0.32	0.18 <sup>†</sup>	0.21	<b>0.42</b>	<b>0.37</b>	0.29	0.24	0.33	<b>0.48</b>	<b>0.48</b>	0.32	0.29	<b>0.44</b>	<b>0.48</b>	0.21
CMU-CMB	<b>0.50</b>		0.26	0.29	<b>0.42</b>	0.4	<b>0.44</b>	<b>0.48</b>	<b>0.49</b>	<b>0.38</b>	<b>0.45</b>	<b>0.55</b>	0.32	0.34	0.34	<b>0.46</b>
CMU-XFR	<b>0.67<sup>†</sup></b>	<b>0.44</b>		<b>0.60<sup>*</sup></b>	<b>0.75<sup>†</sup></b>	<b>0.58</b>	<b>0.73<sup>†</sup></b>	<b>0.62</b>	<b>0.59</b>	<b>0.54</b>	<b>0.77<sup>†</sup></b>	<b>0.48</b>	<b>0.54</b>	<b>0.65<sup>*</sup></b>	<b>0.71<sup>†</sup></b>	<b>0.58</b>
CUED	<b>0.46</b>	<b>0.41</b>	0.20 <sup>*</sup>		<b>0.47</b>	<b>0.56</b>	<b>0.47</b>	<b>0.51<sup>+</sup></b>	<b>0.41</b>	<b>0.54</b>	<b>0.57</b>	<b>0.37</b>	<b>0.43</b>	<b>0.61</b>	0.39	0.15
CUED-C	0.27	0.22	0.08 <sup>†</sup>	0.20		0.31	<b>0.54</b>	<b>0.52<sup>+</sup></b>	0.32	<b>0.52</b>	<b>0.50</b>	0.31	0.40	<b>0.38</b>	0.30	<b>0.52</b>
LIMSI	0.34	0.4	0.29	0.31	<b>0.41</b>		0.23 <sup>*</sup>	<b>0.52</b>	0.38	<b>0.50</b>	<b>0.39</b>	<b>0.49</b>	0.42	0.32	0.26	0.30
LIUM-SYS	<b>0.37</b>	0.32	0.13 <sup>†</sup>	0.39	0.27	<b>0.60<sup>*</sup></b>		0.24	<b>0.44</b>	0.46	<b>0.46</b>	0.33	0.24 <sup>*</sup>	0.25	0.30	0.19
LI-SYS-C	<b>0.40</b>	0.26	0.24	0.20 <sup>*</sup>	0.13 <sup>*</sup>	0.30	0.24		<b>0.44</b>	0.42	0.43	0.35	0.21 <sup>*</sup>	0.30	0.30	0.31
RBMT3	<b>0.46</b>	0.43	0.26	0.38	<b>0.46</b>	<b>0.48</b>	0.39	0.39		0.41	<b>0.44</b>	0.26	0.36	<b>0.50</b>	<b>0.68<sup>*</sup></b>	<b>0.44</b>
RBMT4	0.36	0.33	0.31	0.36	0.39	0.35	<b>0.50</b>	<b>0.45</b>	<b>0.45</b>		<b>0.49</b>	<b>0.40</b>	0.35	<b>0.57</b>	<b>0.51</b>	<b>0.53</b>
RBMT5	0.37	0.33	0.12 <sup>†</sup>	0.32	0.33	0.33	0.39	<b>0.46</b>	0.25	0.22		0.21	0.37	<b>0.44</b>	<b>0.49</b>	<b>0.57</b>
RBMT6	<b>0.50</b>	0.33	0.37	0.34	<b>0.50</b>	0.39	<b>0.44</b>	<b>0.50</b>	<b>0.48</b>	0.37	<b>0.55</b>		<b>0.42</b>	<b>0.48</b>	0.41	0.41
SAAR	<b>0.50</b>	<b>0.46</b>	0.37	0.38	<b>0.44</b>	<b>0.52</b>	<b>0.6<sup>*</sup></b>	<b>0.54<sup>+</sup></b>	<b>0.44</b>	<b>0.53</b>	<b>0.44</b>	0.29		<b>0.34</b>	<b>0.52</b>	<b>0.50</b>
SAAR-C	0.31	<b>0.47</b>	0.23 <sup>*</sup>	0.30	0.24	<b>0.51</b>	<b>0.50</b>	<b>0.47</b>	0.25	0.31	0.33	0.35	0.26		0.47	0.38
UED	0.35	<b>0.37</b>	0.13 <sup>†</sup>	0.39	<b>0.55</b>	<b>0.50</b>	<b>0.50</b>	<b>0.43</b>	0.24 <sup>*</sup>	0.37	0.36	0.41	0.31	0.47		0.36
UED-CMB	<b>0.57</b>	0.36	0.16	<b>0.46</b>	0.38	0.30	<b>0.63</b>	<b>0.39</b>	0.39	0.37	0.35	<b>0.53</b>	0.27	<b>0.48</b>	0.36	
> OTHERS	0.43	0.37	0.22	0.34	0.41	0.44	0.45	0.45	0.4	0.42	<b>0.47</b>	0.37	0.34	0.43	0.44	0.42
≥ OTHERS	0.66	0.59	0.38	0.55	0.64	0.63	0.66	<b>0.69</b>	0.58	0.58	0.65	0.57	0.54	0.64	0.61	0.61

Table 21: Sentence-level ranking for the French-English News Task.

	CMU-XFR	CUED	DCU	LIMSI	LIUM-SYS	RBMT3	RBMT4	RBMT5	RBMT6	SAAR	SAAR-C	SYSTRAN	UCL	UEDIN
CMU-XFR		<b>0.53</b>	<b>0.50</b>	<b>0.74<sup>†</sup></b>	<b>0.79<sup>*</sup></b>	<b>0.55</b>	0.46	<b>0.50</b>	0.36	<b>0.73</b>	<b>0.92<sup>†</sup></b>	0.36	<b>0.44</b>	<b>0.77<sup>*</sup></b>
CUED	0.29		<b>0.42</b>	<b>0.29</b>	<b>0.48</b>	0.16 <sup>†</sup>	<b>0.53</b>	0.16 <sup>†</sup>	0.18 <sup>†</sup>	0.18 <sup>*</sup>	<b>0.55</b>	0.06 <sup>†</sup>	0.21	0.38
DCU	0.46	0.29		0.38	<b>0.47</b>	0.37	0.27	0.24	0.29	0.35	<b>0.55</b>	0.18	0.25	<b>0.50</b>
LIMSI	0.11 <sup>†</sup>	0.21	<b>0.44</b>		0.11	0.12 <sup>†</sup>	0.17	0.29	0.05 <sup>†</sup>	0.30	0.32	0.19	0.29	0.33
LIUM-SYS	0.14 <sup>*</sup>	0.16	0.24	<b>0.32</b>		0.06 <sup>†</sup>	0.13	0.22	0.12 <sup>†</sup>	0.14 <sup>†</sup>	0.33	0.20 <sup>*</sup>	0.26	0.32
RBMT3	0.36	<b>0.79<sup>†</sup></b>	<b>0.58</b>	<b>0.88<sup>†</sup></b>	<b>0.72<sup>†</sup></b>		0.40	<b>0.57</b>	0.21	<b>0.67</b>	<b>0.72<sup>†</sup></b>	<b>0.50</b>	<b>0.54</b>	<b>0.67</b>
RBMT4	<b>0.50</b>	0.40	<b>0.64</b>	<b>0.67</b>	<b>0.56</b>	0.40		0.42	0.21 <sup>†</sup>	<b>0.52</b>	<b>0.67</b>	0.33	<b>0.47</b>	<b>0.75</b>
RBMT5	0.38	<b>0.79<sup>†</sup></b>	<b>0.60</b>	<b>0.57</b>	<b>0.56</b>	0.24	0.42		0.26	0.48	<b>0.72<sup>†</sup></b>	<b>0.50</b>	<b>0.46</b>	<b>0.60</b>
RBMT6	<b>0.54</b>	<b>0.79<sup>†</sup></b>	<b>0.67</b>	<b>0.77<sup>†</sup></b>	<b>0.82<sup>†</sup></b>	<b>0.47</b>	<b>0.79<sup>†</sup></b>	<b>0.53</b>		<b>0.71<sup>*</sup></b>	<b>0.83<sup>†</sup></b>	<b>0.56</b>	<b>0.47</b>	<b>0.77<sup>†</sup></b>
SAAR	0.27	<b>0.59<sup>*</sup></b>	<b>0.57</b>	<b>0.47</b>	<b>0.71<sup>†</sup></b>	0.22	0.29	0.48	0.18 <sup>*</sup>		<b>0.50</b>	0.35	0.23	<b>0.50</b>
SAAR-C	0.04 <sup>†</sup>	0.15	0.31	<b>0.39</b>	<b>0.48</b>	0.14 <sup>†</sup>	0.24	0.21 <sup>†</sup>	0.08 <sup>†</sup>	0.21		0.17 <sup>†</sup>	0.20	<b>0.57</b>
SYSTRAN	<b>0.50</b>	<b>0.81<sup>†</sup></b>	<b>0.65</b>	<b>0.52</b>	<b>0.64<sup>*</sup></b>	0.38	<b>0.62</b>	0.33	0.32	<b>0.41</b>	<b>0.71<sup>†</sup></b>		<b>0.56</b>	<b>0.55</b>
UCL	0.31	<b>0.64</b>	<b>0.56</b>	<b>0.57</b>	<b>0.47</b>	0.46	0.40	0.39	0.27	<b>0.55</b>	<b>0.60</b>	0.44		<b>0.47</b>
UED	0.24 <sup>*</sup>	<b>0.43</b>	0.35	0.33	<b>0.42</b>	0.28	0.25	0.33	0.15 <sup>†</sup>	0.29	0.26	0.25	0.27	
> OTHERS	0.32	0.50	0.5	0.54	0.55	0.28	0.4	0.35	0.21	0.41	<b>0.59</b>	0.32	0.35	0.55
≥ OTHERS	0.42	0.7	0.64	0.78	<b>0.79</b>	0.40	0.50	0.48	0.32	0.58	0.75	0.47	0.52	0.71

Table 22: Sentence-level ranking for the French-English Europarl Task.

	LIMSI	LIUM-SYS	RBMT3	RBMT4	RBMT5	RBMT6	SAAR	SAAR-C	UEDIN	XEROX
LIMSI		0.29	0.25	<b>0.60</b> <sup>†</sup>	<b>0.52</b>	<b>0.48</b>	0.13	0.30	0.13*	0.17*
LIUM-SYSTRAN	<b>0.36</b>		<b>0.41</b>	<b>0.51</b>	<b>0.41</b>	<b>0.53</b>	0.22	0.26	0.27	0.04 <sup>†</sup>
RBMT3	<b>0.56</b>	0.34		<b>0.48</b>	<b>0.52</b>	<b>0.40</b>	0.31	<b>0.53</b>	0.37	0.11 <sup>†</sup>
RBMT4	0.13 <sup>†</sup>	0.36	0.31		0.29	0.19*	0.26	0.15 <sup>†</sup>	0.17 <sup>†</sup>	0.09 <sup>†</sup>
RBMT5	0.33	0.35	0.29	<b>0.42</b>		0.26	0.17 <sup>†</sup>	0.32	0.17 <sup>†</sup>	0.12 <sup>†</sup>
RBMT6	0.42	0.38	0.37	<b>0.43</b> *	<b>0.44</b>		0.32	0.32	0.28	0.11 <sup>†</sup>
SAAR	<b>0.56</b>	<b>0.52</b>	<b>0.51</b>	<b>0.56</b>	<b>0.69</b> <sup>†</sup>	<b>0.41</b>		<b>0.33</b>	<b>0.46</b>	0.3
SAAR-CONTRAST	<b>0.55</b>	<b>0.44</b>	0.33	<b>0.63</b> <sup>†</sup>	<b>0.56</b>	<b>0.46</b>	0.21		0.41	0.22*
UEDIN	<b>0.48</b> *	<b>0.48</b>	<b>0.41</b>	<b>0.60</b> <sup>†</sup>	<b>0.65</b> <sup>†</sup>	<b>0.53</b>	0.41	<b>0.43</b>		0.09 <sup>†</sup>
XEROX	<b>0.63</b> *	<b>0.74</b> <sup>†</sup>	<b>0.78</b> <sup>†</sup>	<b>0.74</b> <sup>†</sup>	<b>0.71</b> <sup>†</sup>	<b>0.75</b> <sup>†</sup>	<b>0.44</b>	<b>0.64</b> *	<b>0.63</b> <sup>†</sup>	
> OTHERS	0.44	0.43	0.41	<b>0.54</b>	0.53	0.43	0.28	0.37	0.32	0.13
≥ OTHERS	0.67	0.66	0.60	<b>0.78</b>	0.73	0.66	0.51	0.57	0.55	0.32

Table 23: Sentence-level ranking for the English-French News Task.

	LIMSI	LIUM-SYS	RBMT3	RBMT4	RBMT5	RBMT6	SAAR	SAAR-C	UCL	UEDIN
LIMSI		<b>0.23</b>	0.21*	0.32	0.10 <sup>†</sup>	0.15 <sup>†</sup>	0.35	0.27	0.15 <sup>†</sup>	0.17
LIUM-SYSTRAN			0.28	<b>0.39</b>	0.11 <sup>†</sup>	0.21*	0.22	0.40	0.19 <sup>†</sup>	0.15
RBMT3	<b>0.75</b> *	<b>0.59</b>		<b>0.38</b>	0.39	<b>0.49</b>	<b>0.70</b> <sup>†</sup>	<b>0.81</b> <sup>†</sup>	<b>0.47</b>	<b>0.81</b> <sup>†</sup>
RBMT4	<b>0.64</b>	0.36	0.28		0.24*	0.18	<b>0.61</b>	<b>0.48</b>	0.42	<b>0.50</b>
RBMT5	<b>0.85</b> <sup>†</sup>	<b>0.89</b> <sup>†</sup>	<b>0.49</b>	<b>0.62</b> *		<b>0.67</b> *	<b>0.78</b> <sup>†</sup>	<b>0.91</b> <sup>†</sup>	<b>0.63</b> *	<b>0.93</b> <sup>†</sup>
RBMT6	<b>0.85</b> <sup>†</sup>	<b>0.62</b> *	0.26	<b>0.42</b>	0.24*		<b>0.83</b> <sup>†</sup>	<b>0.82</b> <sup>†</sup>	<b>0.47</b>	<b>0.68</b> <sup>†</sup>
SAAR	<b>0.41</b>	<b>0.52</b>	0.17 <sup>†</sup>	0.30	0.11 <sup>†</sup>	0.06 <sup>†</sup>		<b>0.41</b>	0.11 <sup>†</sup>	<b>0.41</b>
SAAR-CONTRAST	<b>0.47</b>	0.40	0.11 <sup>†</sup>	0.26	0.03 <sup>†</sup>	0.06 <sup>†</sup>	0.32		0.27	0.26
UCL	<b>0.80</b> <sup>†</sup>	<b>0.70</b> <sup>†</sup>	0.42	<b>0.47</b>	0.22*	0.44	<b>0.71</b> <sup>†</sup>	<b>0.61</b>		<b>0.78</b> <sup>†</sup>
UEDIN	<b>0.46</b>	<b>0.41</b>	0.11 <sup>†</sup>	0.33	0.04 <sup>†</sup>	0.15 <sup>†</sup>	0.32	<b>0.36</b>	0.03 <sup>†</sup>	
> OTHERS	<b>0.62</b>	0.54	0.26	0.4	0.17	0.27	0.56	0.6	0.32	0.54
≥ OTHERS	<b>0.79</b>	0.78	0.42	0.61	0.26	0.44	0.74	0.79	0.44	0.77

Table 24: Sentence-level ranking for the English-French Europarl Task.

	BBN-CMB	CMU-XFR	LIMSI	LIU	RBMT2	RBMT3	RBMT4	RBMT5	RBMT6	SAAR	SAAR-C	UEDIN	UEDIN-CMB
BBN-COMBO		0.1 <sup>†</sup>	0.22	0.37	<b>0.62</b> *	<b>0.69</b> <sup>†</sup>	<b>0.74</b> *	<b>0.66</b> <sup>†</sup>	0.41	<b>0.63</b> *	<b>0.60</b> *	<b>0.35</b>	<b>0.40</b>
CMU-STATXFER	<b>0.71</b> <sup>†</sup>		<b>0.44</b>	<b>0.54</b>	<b>0.76</b> <sup>†</sup>	<b>0.79</b> <sup>†</sup>	<b>0.73</b> <sup>†</sup>	<b>0.74</b> <sup>†</sup>	<b>0.80</b> <sup>†</sup>	<b>0.62</b> <sup>†</sup>	<b>0.65</b> <sup>†</sup>	<b>0.54</b> <sup>†</sup>	0.37
LIMSI	<b>0.44</b>	0.24		<b>0.41</b>	<b>0.67</b> *	<b>0.65</b> <sup>†</sup>	<b>0.69</b> <sup>†</sup>	<b>0.54</b>	<b>0.50</b>	<b>0.50</b>	<b>0.63</b>	0.38	0.22
LIU	0.37	0.27	0.34		<b>0.55</b> *	<b>0.56</b>	<b>0.61</b> <sup>†</sup>	<b>0.50</b>	<b>0.45</b>	<b>0.48</b>	<b>0.56</b>	<b>0.32</b>	0.34
RBMT2	0.21*	0.14 <sup>†</sup>	0.31*	0.20*		0.27	0.43	0.29	0.34	0.30	0.13 <sup>†</sup>	0.25 <sup>†</sup>	0.24*
RBMT3	0.18 <sup>†</sup>	0.13 <sup>†</sup>	0.19 <sup>†</sup>	0.27	<b>0.56</b>		0.37	0.33	0.32	0.29	0.29	0.19 <sup>†</sup>	0.17 <sup>†</sup>
RBMT4	0.22*	0.12 <sup>†</sup>	0.17 <sup>†</sup>	0.18 <sup>†</sup>	<b>0.46</b>	<b>0.51</b>		0.3	0.31	0.18 <sup>†</sup>	0.26*	0.28	0.17 <sup>†</sup>
RBMT5	0.22 <sup>†</sup>	0.12 <sup>†</sup>	0.32	0.36	<b>0.58</b>	<b>0.51</b>	<b>0.40</b>		0.29	0.23*	0.37	0.3	0.28
RBMT6	<b>0.55</b>	0.08 <sup>†</sup>	0.40	0.4	<b>0.51</b>	<b>0.51</b>	<b>0.47</b>	<b>0.51</b>		<b>0.49</b>	<b>0.52</b>	0.22*	0.43
SAAR	0.23*	0.21 <sup>†</sup>	0.40	0.39	<b>0.52</b>	<b>0.50</b>	<b>0.61</b> <sup>†</sup>	<b>0.53</b> *	0.38		<b>0.50</b> *	0.26*	0.13*
SAAR-CONTRAST	0.23*	0.19 <sup>†</sup>	0.3	0.37	<b>0.71</b> <sup>†</sup>	<b>0.37</b>	<b>0.60</b> *	0.37	0.33	0.17*		<b>0.48</b>	0.13*
UEDIN	0.23	0.13 <sup>†</sup>	0.38	0.3	<b>0.68</b> <sup>†</sup>	<b>0.65</b> <sup>†</sup>	<b>0.55</b>	<b>0.59</b>	<b>0.64</b> *	<b>0.67</b> *	0.38		<b>0.42</b>
UEDIN-COMBO	0.35	<b>0.41</b>	<b>0.59</b>	<b>0.50</b>	<b>0.72</b> *	<b>0.66</b> <sup>†</sup>	<b>0.83</b> <sup>†</sup>	<b>0.56</b>	<b>0.52</b>	<b>0.50</b> *	<b>0.67</b> *	0.38	
> OTHERS	0.32	0.17	0.34	0.35	<b>0.61</b>	0.56	0.57	0.49	0.45	0.41	0.46	0.33	0.28
≥ OTHERS	0.51	0.35	0.52	0.56	<b>0.74</b>	0.73	0.73	0.67	0.59	0.61	0.65	0.55	0.44

Table 25: Sentence-level ranking for the German-English News Task.

	CMU-XFR	LIMSI	LIU	RBMT2	RBMT3	RBMT4	RBMT5	RBMT6	SAAR	UCL	UEDIN
CMU-STATXFER		<b>0.57*</b>	<b>0.77<sup>†</sup></b>	<b>0.53</b>	<b>0.71<sup>†</sup></b>	<b>0.69<sup>†</sup></b>	<b>0.50</b>	<b>0.58</b>	<b>0.82<sup>†</sup></b>	<b>0.46</b>	<b>0.75<sup>†</sup></b>
LIMSI	0.17*		0.35	<b>0.71*</b>	<b>0.63</b>	<b>0.76*</b>	<b>0.50</b>	<b>0.59</b>	<b>0.52</b>	0.23	<b>0.67<sup>†</sup></b>
LIU	0.14 <sup>†</sup>	0.35		<b>0.50</b>	0.29	<b>0.67</b>	0.3	0.42	<b>0.35</b>	0.27	<b>0.57</b>
RBMT2	0.27	0.24*	0.46		0.39	0.33	0.36	0.42	<b>0.50</b>	0.33	<b>0.46</b>
RBMT3	0.23 <sup>†</sup>	0.3	<b>0.57</b>	<b>0.45</b>		0.40	0.31	0.38	<b>0.56</b>	0.32	<b>0.55</b>
RBMT4	0.22 <sup>†</sup>	0.19*	0.29	<b>0.50</b>	<b>0.48</b>		0.39	<b>0.48</b>	0.41	0.32	<b>0.61</b>
RBMT5	0.40	0.40	<b>0.56</b>	<b>0.54</b>	<b>0.57</b>	<b>0.52</b>		0.3	<b>0.48</b>	0.29*	<b>0.54</b>
RBMT6	0.27	0.32	<b>0.48</b>	<b>0.46</b>	<b>0.53</b>	0.44	<b>0.51</b>		<b>0.55</b>	0.36	<b>0.61</b>
SAAR	0.12 <sup>†</sup>	0.19	0.30	0.44	0.41	<b>0.48</b>	0.32	0.42		0.20 <sup>†</sup>	0.40
UCL	0.35	<b>0.54</b>	<b>0.46</b>	<b>0.63</b>	<b>0.61</b>	<b>0.68</b>	<b>0.68*</b>	<b>0.61</b>	<b>0.63<sup>†</sup></b>		<b>0.65<sup>†</sup></b>
UEDIN	0.22 <sup>†</sup>	0.17 <sup>†</sup>	0.32	0.42	0.42	0.36	0.41	0.27	0.40	0.23 <sup>†</sup>	
> OTHERS	0.24	0.32	0.46	0.51	0.51	0.53	0.43	0.43	0.53	0.30	<b>0.58</b>
≥ OTHERS	0.36	0.49	0.61	0.63	0.6	0.61	0.54	0.54	0.68	0.42	<b>0.68</b>

Table 26: Sentence-level ranking for the German-English Europarl Task.

	LIMSI	LIU	RBMT2	RBMT3	RBMT4	RBMT5	RBMT6	SAAR	UEDIN
LIMSI		<b>0.44</b>	<b>0.8<sup>†</sup></b>	<b>0.67<sup>†</sup></b>	<b>0.81<sup>†</sup></b>	<b>0.76<sup>†</sup></b>	<b>0.63<sup>†</sup></b>	<b>0.53</b>	<b>0.47*</b>
LIU	0.29		<b>0.80<sup>†</sup></b>	<b>0.68<sup>†</sup></b>	<b>0.81<sup>†</sup></b>	<b>0.62<sup>†</sup></b>	<b>0.63<sup>†</sup></b>	0.25	0.31
RBMT2	0.13 <sup>†</sup>	0.07 <sup>†</sup>		0.35	0.33	0.32*	0.20 <sup>†</sup>	0.17 <sup>†</sup>	0.09 <sup>†</sup>
RBMT3	0.18 <sup>†</sup>	0.27 <sup>†</sup>	<b>0.50</b>		<b>0.52</b>	<b>0.45</b>	0.29 <sup>†</sup>	0.26	0.21 <sup>†</sup>
RBMT4	0.09 <sup>†</sup>	0.12 <sup>†</sup>	<b>0.47</b>	0.30		<b>0.42</b>	0.22 <sup>†</sup>	0.15 <sup>†</sup>	0.17 <sup>†</sup>
RBMT5	0.12 <sup>†</sup>	0.26 <sup>†</sup>	<b>0.59*</b>	0.42	0.40		0.33	0.28	0.24 <sup>†</sup>
RBMT6	0.25 <sup>†</sup>	0.22 <sup>†</sup>	<b>0.6<sup>†</sup></b>	<b>0.61<sup>†</sup></b>	<b>0.63<sup>†</sup></b>	<b>0.50</b>		0.36	0.33
SAAR	0.28	<b>0.63</b>	<b>0.66<sup>†</sup></b>	<b>0.56</b>	<b>0.7<sup>†</sup></b>	<b>0.62</b>	<b>0.46</b>		<b>0.45</b>
UEDIN	0.24*	<b>0.42</b>	<b>0.75<sup>†</sup></b>	<b>0.66<sup>†</sup></b>	<b>0.73<sup>†</sup></b>	<b>0.68<sup>†</sup></b>	<b>0.51</b>	0.36	
> OTHERS	0.19	0.28	<b>0.64</b>	0.54	0.61	0.54	0.40	0.3	0.27
≥ OTHERS	0.36	0.43	<b>0.79</b>	0.66	0.75	0.67	0.56	0.46	0.44

Table 27: Sentence-level ranking for the English-German News Task.

	CMU-GIMPEL	LIMSI	LIU	RBMT2	RBMT3	RBMT4	RBMT5	RBMT6	SAAR	UCL	UEDIN
CMU-GIMPEL		0.29	0.28	0.41	<b>0.49</b>	<b>0.56</b>	0.44	0.24 <sup>†</sup>	0.09*	0.24*	<b>0.52</b>
LIMSI	<b>0.45</b>		0.31	0.48	0.45	<b>0.54</b>	0.40	0.35	0.40	0.29*	<b>0.47</b>
LIU	<b>0.34</b>	<b>0.47</b>		<b>0.56</b>	0.44	<b>0.65*</b>	0.37	0.30	0.31	0.19 <sup>†</sup>	<b>0.50</b>
RBMT2	<b>0.51</b>	0.48	0.41		0.41	<b>0.48</b>	0.22 <sup>†</sup>	0.24*	<b>0.62</b>	0.26*	0.43
RBMT3	0.40	<b>0.50</b>	<b>0.47</b>	<b>0.47</b>		<b>0.60</b>	0.33	0.3*	0.11	0.26*	<b>0.50</b>
RBMT4	0.39	0.37	0.27*	0.41	0.35		0.22 <sup>†</sup>	0.14 <sup>†</sup>	0.25	0.33	0.46
RBMT5	<b>0.49</b>	<b>0.47</b>	<b>0.54</b>	<b>0.64<sup>†</sup></b>	<b>0.60</b>	<b>0.64<sup>†</sup></b>		0.32	0.47	0.45	<b>0.64<sup>†</sup></b>
RBMT6	<b>0.71<sup>†</sup></b>	<b>0.50</b>	<b>0.58</b>	<b>0.57*</b>	<b>0.65*</b>	<b>0.74<sup>†</sup></b>	<b>0.46</b>		<b>0.41</b>	0.36	<b>0.60</b>
SAAR	<b>0.73*</b>	0.40	<b>0.39</b>	0.39	<b>0.78</b>	<b>0.58</b>	0.47	0.35		0.31	<b>0.50</b>
UCL	<b>0.61*</b>	<b>0.6*</b>	<b>0.67<sup>†</sup></b>	<b>0.59*</b>	<b>0.68*</b>	<b>0.64</b>	<b>0.53</b>	<b>0.51</b>	<b>0.62</b>		<b>0.70<sup>†</sup></b>
UEDIN	0.25	0.27	0.30	<b>0.52</b>	0.41	<b>0.49</b>	0.26 <sup>†</sup>	0.31	0.25	0.23 <sup>†</sup>	
> OTHERS	0.47	0.43	0.43	0.51	0.51	<b>0.59</b>	0.36	0.3	0.37	0.3	0.54
≥ OTHERS	0.61	0.58	0.58	0.62	0.58	<b>0.68</b>	0.47	0.43	0.53	0.39	0.67

Table 28: Sentence-level ranking for the English-German Europarl Task.

	CMU-SMT	CUED	CUED-C	LIMSI	RBMT3	RBMT4	RBMT5	RBMT6	SAAR	UCB	UEDIN	UPC
CMU-SMT		<b>0.41</b>	<b>0.62*</b>	0.33	<b>0.54*</b>	<b>0.57<sup>†</sup></b>	<b>0.42</b>	<b>0.46</b>	<b>0.46</b>	0.29	0.34	<b>0.37</b>
CUED	0.29		0.24	0.27	<b>0.54*</b>	<b>0.76<sup>†</sup></b>	<b>0.61*</b>	<b>0.50</b>	<b>0.39</b>	<b>0.46</b>	0.26	<b>0.42</b>
CUED-CONTRAST	0.19*	0.24		0.23	<b>0.47</b>	<b>0.48</b>	0.28	<b>0.41</b>	0.37	0.26	0.26	0.33
LIMSI	0.33	<b>0.30</b>	<b>0.51</b>		0.41	<b>0.56<sup>†</sup></b>	<b>0.47</b>	<b>0.41</b>	<b>0.46</b>	0.33	<b>0.37</b>	<b>0.43</b>
RBMT3	0.19*	0.23*	0.37	<b>0.43</b>		<b>0.39</b>	0.28	0.3	0.33	0.39	0.30	<b>0.49</b>
RBMT4	0.19 <sup>†</sup>	0.14 <sup>†</sup>	0.27	0.21 <sup>†</sup>	0.27		0.21 <sup>†</sup>	0.30	0.27	0.17 <sup>†</sup>	0.29*	0.23*
RBMT5	0.37	0.19*	<b>0.56</b>	0.35	<b>0.47</b>	<b>0.57<sup>†</sup></b>		<b>0.56</b>	<b>0.43</b>	0.24*	0.35	<b>0.52</b>
RBMT6	0.41	0.30	0.29	0.39	<b>0.43</b>	<b>0.50</b>	0.25		<b>0.46</b>	0.34	<b>0.44</b>	<b>0.46</b>
SAAR	0.29	0.25	<b>0.43</b>	0.32	<b>0.50</b>	<b>0.42</b>	0.33	0.31		0.2*	0.26	0.3
UCB	0.29	0.36	<b>0.52</b>	<b>0.49</b>	<b>0.46</b>	<b>0.61<sup>†</sup></b>	<b>0.6*</b>	<b>0.41</b>	<b>0.56*</b>		<b>0.39</b>	0.28
UEDIN	<b>0.39</b>	<b>0.37</b>	<b>0.52</b>	0.30	<b>0.50</b>	<b>0.61*</b>	<b>0.58</b>	0.39	<b>0.46</b>	0.24		<b>0.44</b>
UPC	0.26	0.36	<b>0.47</b>	0.35	0.40	<b>0.59*</b>	0.32	0.42	<b>0.46</b>	<b>0.33</b>	0.41	
> OTHERS	0.29	0.28	0.43	0.34	0.45	<b>0.55</b>	0.39	0.40	0.42	0.29	0.34	0.39
≥ OTHERS	0.57	0.56	0.67	0.58	0.67	<b>0.77</b>	0.58	0.61	0.67	0.54	0.56	0.60

Table 29: Sentence-level ranking for the Spanish-English News Task.

	CMU-SMT	CUED	DCU	LIMSI	RBMT3	RBMT4	RBMT5	RBMT6	SAAR	UCL	UEDIN	UPC
CMU-SMT		0.36	<b>0.38</b>	<b>0.37</b>	0.10 <sup>†</sup>	0.20 <sup>†</sup>	0.14 <sup>†</sup>	0.32	0.39	0.22	0.25	0.38
CUED	<b>0.40</b>		0.38	<b>0.53</b>	0.33	0.30	0.30	0.20 <sup>†</sup>	0.32	0.08 <sup>†</sup>	0.36	0.29
DCU	0.34	0.38		<b>0.46</b>	0.32	0.19*	0.26*	0.21*	0.32	0.33	0.25	<b>0.46</b>
LIMSI	0.31	0.30	0.21		0.05 <sup>†</sup>	0.09 <sup>†</sup>	0.15 <sup>†</sup>	0.18*	0.24	0.10 <sup>†</sup>	0.19	<b>0.48</b>
RBMT3	<b>0.83<sup>†</sup></b>	<b>0.62</b>	<b>0.58</b>	<b>0.73<sup>†</sup></b>		<b>0.56</b>	0.25	0.37	<b>0.60<sup>†</sup></b>	0.31	<b>0.66*</b>	<b>0.78<sup>†</sup></b>
RBMT4	<b>0.73<sup>†</sup></b>	<b>0.54</b>	<b>0.76*</b>	<b>0.74<sup>†</sup></b>	0.28		0.38	0.24	<b>0.53</b>	0.29	<b>0.56</b>	<b>0.65*</b>
RBMT5	<b>0.79<sup>†</sup></b>	<b>0.55</b>	<b>0.67*</b>	<b>0.75<sup>†</sup></b>	<b>0.58</b>	<b>0.57</b>		<b>0.59*</b>	<b>0.70<sup>†</sup></b>	<b>0.44</b>	<b>0.71*</b>	<b>0.67</b>
RBMT6	<b>0.52</b>	<b>0.77<sup>†</sup></b>	<b>0.66*</b>	<b>0.68*</b>	<b>0.42</b>	<b>0.49</b>	0.18*		<b>0.55</b>	0.41	<b>0.54</b>	<b>0.71</b>
SAAR	<b>0.43</b>	<b>0.42</b>	<b>0.41</b>	<b>0.47</b>	0.20 <sup>†</sup>	0.32	0.17 <sup>†</sup>	0.30		0.22*	0.35	<b>0.32</b>
UCL	<b>0.56</b>	<b>0.71<sup>†</sup></b>	<b>0.56</b>	<b>0.70<sup>†</sup></b>	<b>0.42</b>	<b>0.57</b>	0.33	<b>0.44</b>	<b>0.59*</b>		<b>0.81<sup>†</sup></b>	<b>0.67</b>
UEDIN	<b>0.28</b>	<b>0.46</b>	<b>0.39</b>	<b>0.31</b>	0.29*	0.42	0.25*	0.39	0.35	0.15 <sup>†</sup>		0.40
UPC	<b>0.44</b>	<b>0.39</b>	0.43	0.36	0.07 <sup>†</sup>	0.23*	0.24	0.29	0.27	0.20	0.40	
> OTHERS	0.50	0.5	0.49	<b>0.53</b>	0.28	0.36	0.24	0.32	0.44	0.26	0.45	0.51
≥ OTHERS	0.71	0.68	0.68	<b>0.78</b>	0.43	0.49	0.35	0.47	0.67	0.43	0.66	0.69

Table 30: Sentence-level ranking for the Spanish-English Europarl Task.

	CMU-SMT	LIMSI	RBMT3	RBMT4	RBMT5	RBMT6	SAAR	UCB	UEDIN	UPC
CMU-SMT		0.39	<b>0.57</b>	<b>0.52*</b>	<b>0.62<sup>†</sup></b>	<b>0.56*</b>	<b>0.50</b>	<b>0.41</b>	<b>0.42</b>	<b>0.56<sup>†</sup></b>
LIMSI	<b>0.42</b>		<b>0.56</b>	<b>0.53</b>	<b>0.63*</b>	<b>0.58</b>	0.32	0.39	0.35	0.35
RBMT3	0.23	0.3		0.34	<b>0.46</b>	<b>0.50</b>	0.39	0.17	0.21 <sup>†</sup>	0.06*
RBMT4	0.25*	0.30	<b>0.47</b>		0.31	0.35	0.38	0.36	0.32	0.19
RBMT5	0.21 <sup>†</sup>	0.20*	0.28	<b>0.42</b>		0.42	0.29*	0.24	0.17 <sup>†</sup>	0.23
RBMT6	0.23*	0.23	0.31	<b>0.41</b>	0.42		0.23*	0.19	0.24*	0.24
SAAR	0.36	<b>0.52</b>	0.39	<b>0.43</b>	<b>0.67*</b>	<b>0.54*</b>		0.36	0.29	0.42
UCB	0.37	0.39	<b>0.52</b>	<b>0.39</b>	<b>0.49</b>	<b>0.52</b>	<b>0.46</b>		0.27	0.25
UEDIN	0.35	<b>0.48</b>	<b>0.62<sup>†</sup></b>	<b>0.48</b>	<b>0.64<sup>†</sup></b>	<b>0.61*</b>	<b>0.50</b>	<b>0.47</b>		<b>0.53*</b>
UPC	0.11 <sup>†</sup>	<b>0.41</b>	<b>0.63*</b>	<b>0.48</b>	<b>0.50</b>	<b>0.57</b>	0.42	<b>0.63</b>	0.06*	
> OTHERS	0.28	0.36	0.47	0.45	<b>0.52</b>	0.51	0.38	0.34	0.27	0.33
≥ OTHERS	0.49	0.54	0.68	0.67	<b>0.72</b>	0.72	0.55	0.59	0.48	0.60

Table 31: Sentence-level ranking for the English-Spanish News Task.

	CMU-SMT	LIMSI	RBMT3	RBMT4	RBMT5	RBMT6	SAAR	UCL	UEDIN	UPC	UW
CMU-SMT		0.28	<b>0.47</b>	0.33	0.17 <sup>†</sup>	0.26	<b>0.50</b>	0.25	<b>0.48*</b>	<b>0.44</b>	0.28
LIMSI	<b>0.38</b>		0.19*	0.33	0.16*	0.23	0.33	0.14 <sup>†</sup>	0.14	0.35	0.32
RBMT3	0.42	<b>0.62*</b>		0.42	0.36	0.29	<b>0.54</b>	0.28	0.39	0.50	<b>0.75<sup>†</sup></b>
RBMT4	<b>0.46</b>	<b>0.47</b>	0.42		0.19	0.31	<b>0.61</b>	<b>0.50</b>	0.40	<b>0.50</b>	<b>0.57</b>
RBMT5	<b>0.70<sup>†</sup></b>	<b>0.64*</b>	<b>0.59</b>	<b>0.48</b>		0.35	<b>0.65*</b>	<b>0.52</b>	<b>0.64</b>	<b>0.61</b>	<b>0.63*</b>
RBMT6	<b>0.63</b>	<b>0.58</b>	<b>0.47</b>	<b>0.56</b>	<b>0.50</b>		<b>0.78<sup>†</sup></b>	0.32	<b>0.58</b>	0.33	<b>0.71*</b>
SAAR	0.33	<b>0.40</b>	0.33	0.30	0.23*	0.19 <sup>†</sup>		0.20	0.27	0.24	0.33
UCL	<b>0.46</b>	<b>0.64<sup>†</sup></b>	<b>0.41</b>	0.46	0.36	<b>0.41</b>	<b>0.60</b>		<b>0.65*</b>	<b>0.42</b>	<b>0.57*</b>
UEDIN	0.09*	<b>0.29</b>	<b>0.48</b>	<b>0.45</b>	0.28	0.27	<b>0.41</b>	0.19*		0.25	0.17
UPC	0.22	<b>0.40</b>	0.50	0.43	0.28	<b>0.40</b>	<b>0.52</b>	0.26	<b>0.56</b>		<b>0.58</b>
UW	<b>0.44</b>	0.32	0.06 <sup>†</sup>	0.29	0.17*	0.21*	0.33	0.14*	<b>0.33</b>	0.33	
> OTHERS	0.43	0.46	0.4	0.4	0.26	0.28	<b>0.53</b>	0.28	0.46	0.4	0.49
≥ OTHERS	0.67	<b>0.74</b>	0.55	0.56	0.41	0.44	0.72	0.50	0.71	0.59	0.74

Table 32: Sentence-level ranking for the English-Spanish Europarl Task.

	DCU	UEDIN	UMD
DCU		0.26 <sup>†</sup>	0.4
UEDIN	<b>0.37<sup>†</sup></b>		<b>0.46<sup>†</sup></b>
UMD	0.4	0.31 <sup>†</sup>	
> OTHERS	0.38	0.28	<b>0.43</b>
≥ OTHERS	<b>0.68</b>	0.58	0.65

Table 33: Sentence-level ranking for the Czech-English News Task.

	DCU	SYSTRAN	UEDIN	UMD
DCU		0.21 <sup>†</sup>	0.19 <sup>†</sup>	0.37
SYSTRAN	<b>0.59<sup>†</sup></b>		<b>0.47<sup>†</sup></b>	<b>0.61<sup>†</sup></b>
UEDIN	<b>0.42<sup>†</sup></b>	0.27 <sup>†</sup>		<b>0.50<sup>†</sup></b>
UMD	<b>0.38</b>	0.18 <sup>†</sup>	0.29 <sup>†</sup>	
> OTHERS	0.46	0.22	0.31	<b>0.49</b>
≥ OTHERS	<b>0.75</b>	0.45	0.60	0.72

Table 34: Sentence-level ranking for the Czech-English Commentary Task.

	CU-BOJAR	CU-TECTOMT	PC-TRANSLATOR	UEDIN
CU-BOJAR		0.32 <sup>†</sup>	<b>0.51<sup>†</sup></b>	0.27 <sup>†</sup>
CU-TECTOMT	<b>0.52<sup>†</sup></b>		<b>0.58<sup>†</sup></b>	<b>0.42</b>
PC-TRANSLATOR	0.35 <sup>†</sup>	0.25 <sup>†</sup>		0.26 <sup>†</sup>
UEDIN	<b>0.5<sup>†</sup></b>	0.40	<b>0.59<sup>†</sup></b>	
> OTHERS	0.45	0.32	<b>0.56</b>	0.32
≥ OTHERS	0.63	0.49	<b>0.72</b>	0.50

Table 35: Sentence-level ranking for the English-Czech News Task.

	CU-BOJAR	CU-TECTOMT	PC-TRANSLATOR	UEDIN
CU-BOJAR		0.28 <sup>†</sup>	0.38	0.19 <sup>†</sup>
CU-TECTOMT	<b>0.58<sup>†</sup></b>		<b>0.53<sup>†</sup></b>	<b>0.43</b>
PC-TRANSLATOR	<b>0.45</b>	0.3 <sup>†</sup>		0.26 <sup>†</sup>
UEDIN	<b>0.60<sup>†</sup></b>	0.37	<b>0.56<sup>†</sup></b>	
> OTHERS	<b>0.54</b>	0.32	0.49	0.29
≥ OTHERS	<b>0.71</b>	0.49	0.66	0.49

Table 36: Sentence-level ranking for the English-Czech Commentary Task.

	MORPHOLOGIC	MLOGIC	UEDIN
	UEDIN	<b>0.68<sup>†</sup></b>	0.15 <sup>†</sup>
>	OTHERS	<b>0.68</b>	0.15
≥	OTHERS	<b>0.85</b>	0.32

Table 37: Sentence-level ranking for the Hungarian-English News Task.

	CMU-XFR	CUED	CUED-C	LIMSI	LIUM-SYS	LIUM-SYS-C	RBMT3	RBMT4	RBMT5	RBMT6	SAAR	SAAR-C	UEDIN
CMU-XFR		<b>0.37</b>	<b>0.49<sup>†</sup></b>	<b>0.62<sup>†</sup></b>	<b>0.57<sup>†</sup></b>	<b>0.61<sup>†</sup></b>	<b>0.49</b>	<b>0.49</b>	<b>0.48*</b>	<b>0.41</b>	<b>0.56<sup>†</sup></b>	<b>0.39</b>	<b>0.46*</b>
CUED	0.28		<b>0.21</b>	<b>0.30</b>	<b>0.30</b>	0.13	0.28	0.18	<b>0.27</b>	0.28	0.31	<b>0.34</b>	<b>0.18</b>
CUED-C	0.2 <sup>†</sup>	0.11		<b>0.30*</b>	0.19	<b>0.33</b>	0.18 <sup>†</sup>	0.21	0.24	0.2 <sup>†</sup>	0.2*	0.17*	<b>0.24</b>
LIMSI	0.13 <sup>†</sup>	0.20	0.13*		<b>0.27</b>	0.22	0.23	0.24	0.2	0.20*	0.16*	0.23	0.22
LIUM-SYS	0.18 <sup>†</sup>	0.17	<b>0.27</b>	0.17		<b>0.20</b>	0.18*	<b>0.41</b>	<b>0.29</b>	0.24	0.26	0.22	0.26
LI-SYS-C	0.18 <sup>†</sup>	<b>0.28</b>	0.24	<b>0.25</b>	0.07		0.33	0.2*	0.27	0.18 <sup>†</sup>	0.23	0.25	0.19
RBMT3	0.28	<b>0.34</b>	<b>0.52<sup>†</sup></b>	<b>0.28</b>	<b>0.40*</b>	<b>0.37</b>		0.27	<b>0.46<sup>†</sup></b>	0.27	0.30	<b>0.39</b>	<b>0.34</b>
RBMT4	0.29	<b>0.40</b>	<b>0.34</b>	<b>0.31</b>	0.39	<b>0.43*</b>	<b>0.33</b>		<b>0.34</b>	<b>0.34</b>	0.27	<b>0.41</b>	<b>0.31</b>
RBMT5	0.22*	0.24	<b>0.34</b>	<b>0.3</b>	0.27	<b>0.43</b>	0.14 <sup>†</sup>	0.24		0.13*	0.32	<b>0.32</b>	<b>0.32</b>
RBMT6	0.3	<b>0.41</b>	<b>0.50<sup>†</sup></b>	<b>0.39*</b>	<b>0.33</b>	<b>0.58<sup>†</sup></b>	<b>0.3</b>	0.33	<b>0.37*</b>		0.33	<b>0.52*</b>	<b>0.37</b>
SAAR	0.27 <sup>†</sup>	<b>0.33</b>	<b>0.43*</b>	<b>0.37*</b>	<b>0.4</b>	<b>0.42</b>	<b>0.41</b>	<b>0.36</b>	0.32	<b>0.41</b>		<b>0.23</b>	<b>0.41</b>
SAAR-C	0.28	0.32	<b>0.38*</b>	<b>0.27</b>	<b>0.27</b>	<b>0.45</b>	0.23	0.21	0.20	0.23*	0.18		0.19
UED	0.19*	0.15	0.20	<b>0.25</b>	<b>0.29</b>	0.19	0.28	0.27	0.19	0.24	0.21	<b>0.26</b>	
>	OTHERS	0.24	0.27	0.33	0.32	<b>0.37</b>	0.29	0.28	0.30	0.27	0.29	0.31	0.29
≥	OTHERS	0.51	0.75	0.79	<b>0.80</b>	0.77	0.78	0.65	0.66	0.73	0.64	0.74	0.77

Table 38: Constituent ranking for the French-English News Task

	CMU-XFR	CUED	DCU	LIMSI	LIUM-SYS	RBMT3	RBMT4	RBMT5	RBMT6	SAAR	SAAR-C	SYSTRAN	UCL	UEDIN
CMU-XFR		<b>0.42<sup>†</sup></b>	<b>0.4<sup>†</sup></b>	<b>0.37*</b>	<b>0.54<sup>†</sup></b>	0.16*	0.21	<b>0.41</b>	0.23	<b>0.49<sup>†</sup></b>	<b>0.42<sup>†</sup></b>	0.34	<b>0.45</b>	<b>0.50<sup>†</sup></b>
CUED	0.03 <sup>†</sup>		<b>0.13</b>	<b>0.08</b>	0.14	0.13 <sup>†</sup>	0.13 <sup>†</sup>	0.08 <sup>†</sup>	0.05 <sup>†</sup>	0.08	0.04	0.15	0.11	0.07
DCU	0.09 <sup>†</sup>	0.08		0.10	0.12	0.06 <sup>†</sup>	0.20	<b>0.31</b>	0.16 <sup>†</sup>	0.14	<b>0.22</b>	0.13	0.10	0.16
LIMSI	0.1*	0.05	<b>0.19</b>		0.05	0.04 <sup>†</sup>	0.08 <sup>†</sup>	0.19	0.11 <sup>†</sup>	0.18	0.09	0.05 <sup>†</sup>	0.05 <sup>†</sup>	
LIUM-SYS	0.03 <sup>†</sup>	0.14	<b>0.19</b>	<b>0.07</b>		0	0.08*	0.03 <sup>†</sup>	0.05 <sup>†</sup>	0.03 <sup>†</sup>	0.09	0.15	0.14	0.08
RBMT3	<b>0.44*</b>	<b>0.61<sup>†</sup></b>	<b>0.50<sup>†</sup></b>	<b>0.58<sup>†</sup></b>	<b>0.56<sup>†</sup></b>		<b>0.41*</b>	<b>0.38</b>	<b>0.32</b>	<b>0.37</b>	<b>0.53<sup>†</sup></b>	<b>0.44</b>	<b>0.50*</b>	<b>0.58<sup>†</sup></b>
RBMT4	<b>0.39</b>	<b>0.44<sup>†</sup></b>	<b>0.43</b>	<b>0.45<sup>†</sup></b>	<b>0.35*</b>	0.12*		<b>0.31</b>	0.23	<b>0.42</b>	<b>0.39</b>	0.33	<b>0.32</b>	<b>0.35</b>
RBMT5	0.19	<b>0.47<sup>†</sup></b>	0.29	<b>0.35</b>	<b>0.37<sup>†</sup></b>	0.18	0.17		0.23	<b>0.35</b>	<b>0.33</b>	0.19	<b>0.46</b>	<b>0.40</b>
RBMT6	<b>0.36</b>	<b>0.65<sup>†</sup></b>	<b>0.54<sup>†</sup></b>	<b>0.48<sup>†</sup></b>	<b>0.55<sup>†</sup></b>	0.26	<b>0.40</b>	<b>0.50</b>		<b>0.50<sup>†</sup></b>	<b>0.52<sup>†</sup></b>	<b>0.47*</b>	<b>0.60<sup>†</sup></b>	<b>0.44</b>
SAAR	0.07 <sup>†</sup>	<b>0.25</b>	<b>0.24</b>	0.18	<b>0.37<sup>†</sup></b>	0.23	0.36	0.23	0.12 <sup>†</sup>		0.12	0.23	0.13	<b>0.37*</b>
SAAR-C	0.09 <sup>†</sup>	<b>0.18</b>	0.12	<b>0.16</b>	<b>0.16</b>	0.09 <sup>†</sup>	0.18	0.2	0.06 <sup>†</sup>	0.12		0.09	0.14	<b>0.15</b>
SYSTRAN	0.34	<b>0.40</b>	<b>0.21</b>	<b>0.38<sup>†</sup></b>	<b>0.23</b>	0.25	<b>0.36</b>	<b>0.22</b>	0.15*	0.23	<b>0.28</b>		0.31	<b>0.30*</b>
UCL	0.25	<b>0.34</b>	<b>0.28</b>	<b>0.31<sup>†</sup></b>	<b>0.19</b>	0.11*	0.24	0.23	0.11 <sup>†</sup>	<b>0.24</b>	<b>0.31</b>	<b>0.34</b>		<b>0.37*</b>
UED	0.10 <sup>†</sup>	<b>0.10</b>	0.16	<b>0.05</b>	0.08	0.03 <sup>†</sup>	0.15	0.14	0.18	0.07*	0.13	0.07*	0.11*	
>	OTHERS	0.2	<b>0.32</b>	0.27	0.28	0.12	0.22	0.25	0.15	0.26	0.27	0.22	0.25	0.28
≥	OTHERS	0.63	0.91	0.85	0.91	<b>0.92</b>	0.52	0.65	0.7	0.78	0.87	0.71	0.74	0.89

Table 39: Constituent ranking for the French-English Europarl Task

	LIMSI	LIUM-SYS	RBMT3	RBMT4	RBMT5	RBMT6	SAAR	SAAR-C	UEDIN	XEROX
LIMSI		<b>0.27</b>	<b>0.43</b>	<b>0.43</b>	0.29	<b>0.53*</b>	0.32	<b>0.37</b>	<b>0.30</b>	0.14 <sup>†</sup>
LIUM-SYSTRAN	0.09		0.33	<b>0.36</b>	0.18	<b>0.35</b>	0.16*	0.25	0.22	0.13 <sup>†</sup>
RBMT3	0.36	0.33		0.22	<b>0.31</b>	0.28	<b>0.4</b>	0.26	0.26*	0.20 <sup>†</sup>
RBMT4	0.25	0.26	<b>0.30</b>		0.23	0.16 <sup>†</sup>	0.28	0.26	0.24	0.13 <sup>†</sup>
RBMT5	<b>0.31</b>	<b>0.33</b>	0.22	<b>0.28</b>		0.17	0.27	0.25	0.23	0.13 <sup>†</sup>
RBMT6	0.26*	0.30	<b>0.31</b>	<b>0.38<sup>†</sup></b>	<b>0.32</b>		<b>0.33</b>	0.36	0.39	0.25*
SAAR	0.32	<b>0.41*</b>	0.35	<b>0.38</b>	<b>0.32</b>	0.28		<b>0.14</b>	0.23	0.11 <sup>†</sup>
SAAR-CONTRAST	0.25	<b>0.26</b>	<b>0.36</b>	<b>0.30</b>	<b>0.33</b>	0.36	0.05		0.22	0.13 <sup>†</sup>
UEDIN	0.29	<b>0.34</b>	<b>0.45*</b>	<b>0.4</b>	<b>0.33</b>	<b>0.40</b>	<b>0.31</b>	<b>0.35</b>		0.13 <sup>†</sup>
XEROX	<b>0.66<sup>†</sup></b>	<b>0.55<sup>†</sup></b>	<b>0.61<sup>†</sup></b>	<b>0.65<sup>†</sup></b>	<b>0.58<sup>†</sup></b>	<b>0.51*</b>	<b>0.53<sup>†</sup></b>	<b>0.57<sup>†</sup></b>	<b>0.45<sup>†</sup></b>	
> OTHERS	0.31	0.34	0.38	<b>0.38</b>	0.33	0.33	0.3	0.31	0.29	0.15
≥ OTHERS	0.65	0.76	0.72	<b>0.77</b>	0.76	0.67	0.73	0.75	0.66	0.44

Table 40: Constituent ranking for the English-French News Task

	LIMSI	LIUM-SYS	RBMT3	RBMT4	RBMT5	RBMT6	SAAR	UCL	UEDIN
LIMSI		<b>0.14</b>	0.09 <sup>†</sup>	0.10 <sup>†</sup>	0.24	0.11 <sup>†</sup>	0.13	0.08 <sup>†</sup>	0.12
LIUM-SYSTRAN			0.19 <sup>†</sup>	0.19*	0.15	0.12 <sup>†</sup>	0.06	0.06 <sup>†</sup>	0.09
RBMT3	<b>0.65<sup>†</sup></b>	<b>0.59<sup>†</sup></b>		<b>0.33</b>	<b>0.43</b>	0.32	<b>0.50*</b>	<b>0.39</b>	<b>0.46<sup>†</sup></b>
RBMT4	<b>0.53<sup>†</sup></b>	<b>0.47*</b>	0.19		0.27	0.18*	<b>0.33</b>	<b>0.38</b>	<b>0.39</b>
RBMT5	<b>0.48</b>	<b>0.38</b>	0.32	<b>0.48</b>		<b>0.47</b>	<b>0.55<sup>†</sup></b>	<b>0.44</b>	<b>0.51<sup>†</sup></b>
RBMT6	<b>0.54<sup>†</sup></b>	<b>0.49<sup>†</sup></b>	0.32	<b>0.41*</b>	0.26		<b>0.52<sup>†</sup></b>	<b>0.45</b>	<b>0.58<sup>†</sup></b>
SAAR	<b>0.21</b>	<b>0.17</b>	0.23*	0.25	0.21 <sup>†</sup>	0.17 <sup>†</sup>		0.19	<b>0.13</b>
UCL	<b>0.37<sup>†</sup></b>	<b>0.33<sup>†</sup></b>	0.38	0.35	0.36	0.32	<b>0.34</b>		<b>0.31<sup>†</sup></b>
UEDIN	0.12	<b>0.11</b>	0.17 <sup>†</sup>	0.23	0.13 <sup>†</sup>	0.13 <sup>†</sup>	0.07	0.07 <sup>†</sup>	
> OTHERS	<b>0.38</b>	0.36	0.25	0.30	0.26	0.24	0.33	0.27	0.34
≥ OTHERS	0.88	<b>0.88</b>	0.56	0.68	0.55	0.56	0.81	0.66	0.87

Table 41: Constituent ranking for the English-French Europarl Task

	CMU-XFER	LIMSI	LIU	RBMT2	RBMT3	RBMT4	RBMT5	RBMT6	SAAR	SAAR-C	UEDIN
CMU-STATXFER		<b>0.47<sup>†</sup></b>	<b>0.44</b>	<b>0.52<sup>†</sup></b>	<b>0.53<sup>†</sup></b>	<b>0.57<sup>†</sup></b>	<b>0.49*</b>	<b>0.41</b>	<b>0.49</b>	<b>0.58<sup>†</sup></b>	<b>0.49<sup>†</sup></b>
LIMSI	0.17 <sup>†</sup>		0.18	<b>0.35</b>	0.34	<b>0.40</b>	<b>0.33</b>	<b>0.43</b>	0.19	0.28	0.19
LIU	0.25	<b>0.3</b>		<b>0.37</b>	<b>0.35</b>	<b>0.44</b>	0.28	<b>0.40</b>	0.21	<b>0.33</b>	<b>0.32*</b>
RBMT2	0.19 <sup>†</sup>	0.26	0.30		0.19	<b>0.32</b>	0.16*	0.20	0.26	<b>0.23</b>	0.21
RBMT3	0.22 <sup>†</sup>	<b>0.36</b>	0.26	<b>0.23</b>		0.24	0.23	0.14 <sup>†</sup>	<b>0.15</b>	<b>0.28</b>	0.29
RBMT4	0.20 <sup>†</sup>	0.35	0.23	0.21	0.24		0.22	0.19*	<b>0.36</b>	<b>0.32</b>	<b>0.31</b>
RBMT5	0.26*	0.28	<b>0.38</b>	<b>0.34*</b>	<b>0.31</b>	<b>0.35</b>		0.26	<b>0.3</b>	<b>0.43<sup>†</sup></b>	<b>0.35</b>
RBMT6	0.38	0.37	0.39	<b>0.34</b>	<b>0.44<sup>†</sup></b>	<b>0.4*</b>	<b>0.30</b>		<b>0.28</b>	0.26	<b>0.38</b>
SAAR	0.29	<b>0.22</b>	<b>0.37</b>	<b>0.29</b>	0.10	0.28	0.19	0.22		<b>0.26</b>	0.18
SAAR-CONTRAST	0.18 <sup>†</sup>	<b>0.33</b>	0.29	0.19	0.22	0.24	0.15 <sup>†</sup>	0.26	0.18		0.23
UEDIN	0.11 <sup>†</sup>	<b>0.3</b>	0.13*	<b>0.23</b>	<b>0.35</b>	0.3	0.2	0.37	<b>0.30</b>	<b>0.31</b>	
> OTHERS	0.22	0.33	0.3	0.31	0.32	<b>0.35</b>	0.25	0.29	0.28	0.33	0.30
≥ OTHERS	0.50	0.72	0.67	0.77	0.76	0.74	0.67	0.64	0.76	<b>0.78</b>	0.74

Table 42: Constituent ranking for the German-English News Task



	CMU-XFR	LIMSI	LIU	RBMT2	RBMT3	RBMT4	RBMT5	RBMT6	SAAR	UCL	UEDIN
CMU-STATXFER		<b>0.51</b> <sup>†</sup>	<b>0.51</b> <sup>†</sup>	0.38	<b>0.38</b>	<b>0.41</b>	0.37	<b>0.44</b>	<b>0.48</b> <sup>†</sup>	<b>0.39</b>	<b>0.6</b> <sup>†</sup>
LIMSI	0.18 <sup>†</sup>		0.22	0.3	0.30	0.23	0.22 <sup>†</sup>	0.32	<b>0.27</b>	0.18*	<b>0.29</b>
LIU	0.14 <sup>†</sup>	0.22		0.26*	0.32	0.22*	0.16 <sup>†</sup>	0.31	<b>0.20</b>	0.08 <sup>†</sup>	0.12
RBMT2	0.38	<b>0.51</b>	<b>0.52</b> *		<b>0.40</b>	<b>0.32</b>	0.25	0.31	<b>0.51</b>	0.40	<b>0.7</b> <sup>†</sup>
RBMT3	0.32	<b>0.42</b>	<b>0.45</b>	0.28		<b>0.46</b>	0.16	0.20*	<b>0.56</b> <sup>†</sup>	0.38	<b>0.43</b>
RBMT4	0.32	<b>0.45</b>	<b>0.52</b> *	0.31	0.24		0.13 <sup>†</sup>	0.30	<b>0.49</b> <sup>†</sup>	<b>0.44</b>	<b>0.48</b> *
RBMT5	<b>0.44</b>	<b>0.57</b> <sup>†</sup>	<b>0.53</b> <sup>†</sup>	<b>0.34</b>	<b>0.31</b>	<b>0.43</b> <sup>†</sup>		0.19	<b>0.54</b> <sup>†</sup>	0.39	<b>0.54</b> <sup>†</sup>
RBMT6	0.33	<b>0.51</b>	<b>0.48</b>	<b>0.33</b>	<b>0.47</b> *	<b>0.33</b>	<b>0.33</b>		<b>0.47</b> *	<b>0.42</b>	<b>0.51</b> *
SAAR	0.12 <sup>†</sup>	0.1	0.15	0.26	0.09 <sup>†</sup>	0.19 <sup>†</sup>	0.17 <sup>†</sup>	0.23*		0.11 <sup>†</sup>	0.14
UCL	0.30	<b>0.43</b> *	<b>0.49</b> <sup>†</sup>	0.40	<b>0.40</b>	0.30	<b>0.41</b>	0.39	<b>0.38</b> <sup>†</sup>		<b>0.51</b> <sup>†</sup>
UEDIN	0.11 <sup>†</sup>	0.16	0.12	0.18 <sup>†</sup>	0.25	0.2*	0.18 <sup>†</sup>	0.23*	0.14	0.12 <sup>†</sup>	
> OTHERS	0.27	0.40	0.41	0.31	0.32	0.32	0.25	0.3	0.41	0.30	<b>0.44</b>
≥ OTHERS	0.55	0.75	0.8	0.58	0.64	0.64	0.58	0.59	<b>0.84</b>	0.60	0.83

Table 43: Constituent ranking for the German-English Europarl Task

	LIMSI	LIU	RBMT2	RBMT3	RBMT4	RBMT5	RBMT6	SAAR	UEDIN
LIMSI		0.29	<b>0.46</b>	<b>0.45</b>	<b>0.37</b>	0.36	0.29 <sup>†</sup>	0.33	0.22
LIU	<b>0.32</b>		<b>0.53</b> <sup>†</sup>	<b>0.45</b> *	<b>0.51</b> <sup>†</sup>	<b>0.5</b> *	0.38	<b>0.31</b>	<b>0.36</b>
RBMT2	0.33	0.32 <sup>†</sup>		0.29	0.29	0.20 <sup>†</sup>	0.25 <sup>†</sup>	0.28	0.28 <sup>†</sup>
RBMT3	0.34	0.3*	<b>0.4</b>		<b>0.33</b>	0.3*	0.34	0.20*	0.27 <sup>†</sup>
RBMT4	0.26	0.25 <sup>†</sup>	<b>0.31</b>	0.3		0.23*	0.23 <sup>†</sup>	0.20*	0.21 <sup>†</sup>
RBMT5	<b>0.46</b>	0.33*	<b>0.55</b> <sup>†</sup>	<b>0.46</b> *	<b>0.40</b> *		0.32	0.32	0.29 <sup>†</sup>
RBMT6	<b>0.52</b> <sup>†</sup>	<b>0.40</b>	<b>0.47</b> <sup>†</sup>	<b>0.44</b>	<b>0.53</b> <sup>†</sup>	<b>0.40</b>		0.27	0.37
SAAR	<b>0.38</b>	0.3	<b>0.39</b>	<b>0.42</b> *	<b>0.44</b> *	<b>0.40</b>	<b>0.44</b>		0.34
UEDIN	<b>0.30</b>	0.24	<b>0.53</b> <sup>†</sup>	<b>0.52</b> <sup>†</sup>	<b>0.51</b> <sup>†</sup>	<b>0.56</b> <sup>†</sup>	<b>0.45</b>	<b>0.36</b>	
> OTHERS	0.36	0.31	<b>0.46</b>	0.41	0.42	0.37	0.33	0.28	0.29
≥ OTHERS	0.65	0.57	0.72	0.68	<b>0.75</b>	0.60	0.56	0.61	0.56

Table 44: Constituent ranking for the English-German News Task

	CMU-GIMPEL	LIMSI	LIU	RBMT2	RBMT3	RBMT4	RBMT5	RBMT6	SAAR	UCL	UEDIN
CMU-GIMPEL		0.12	<b>0.27</b>	0.21 <sup>†</sup>	0.30	0.21 <sup>†</sup>	0.27*	0.21 <sup>†</sup>	0.22	0.22	<b>0.23</b>
LIMSI	<b>0.22</b>		<b>0.22</b>	0.34	0.29*	0.29 <sup>†</sup>	0.23 <sup>†</sup>	0.29 <sup>†</sup>	0.2	0.21	<b>0.19</b>
LIU	0.18	0.2		0.20 <sup>†</sup>	0.25*	0.17 <sup>†</sup>	0.16 <sup>†</sup>	0.12 <sup>†</sup>	<b>0.28</b>	0.21	0.18
RBMT2	<b>0.54</b> <sup>†</sup>	<b>0.41</b>	<b>0.62</b> <sup>†</sup>		0.28	0.33	0.35	0.28	<b>0.61</b> *	<b>0.43</b>	<b>0.47</b> <sup>†</sup>
RBMT3	<b>0.47</b>	<b>0.47</b> *	<b>0.47</b> *	<b>0.4</b>		<b>0.33</b>	0.32	0.28	<b>0.56</b> *	<b>0.47</b>	<b>0.48</b> <sup>†</sup>
RBMT4	<b>0.52</b> <sup>†</sup>	<b>0.57</b> <sup>†</sup>	<b>0.52</b> <sup>†</sup>	<b>0.42</b>	0.32		0.27*	0.28	<b>0.47</b>	<b>0.45</b>	<b>0.39</b>
RBMT5	<b>0.49</b> *	<b>0.57</b> <sup>†</sup>	<b>0.65</b> <sup>†</sup>	<b>0.42</b>	<b>0.38</b>	<b>0.48</b> *		0.31	<b>0.76</b> <sup>†</sup>	<b>0.51</b>	<b>0.52</b> <sup>†</sup>
RBMT6	<b>0.51</b> <sup>†</sup>	<b>0.54</b> <sup>†</sup>	<b>0.60</b> <sup>†</sup>	<b>0.41</b>	<b>0.39</b>	<b>0.40</b>	<b>0.41</b>		<b>0.51</b> *	<b>0.53</b> *	<b>0.51</b> <sup>†</sup>
SAAR	<b>0.24</b>	<b>0.29</b>	0.17	0.26*	0.22*	0.25	0.20 <sup>†</sup>	0.21*		<b>0.31</b>	0.12
UCL	<b>0.28</b>	<b>0.32</b>	<b>0.29</b>	0.33	0.38	0.32	0.32	0.29*	0.19		<b>0.30</b>
UEDIN	0.1	0.13	<b>0.22</b>	0.2 <sup>†</sup>	0.18 <sup>†</sup>	0.22	0.21 <sup>†</sup>	0.18 <sup>†</sup>	<b>0.15</b>	0.17	
> OTHERS	0.37	0.37	<b>0.42</b>	0.32	0.30	0.31	0.28	0.25	0.39	0.35	0.35
≥ OTHERS	0.77	0.75	0.81	0.58	0.59	0.58	0.51	0.52	0.77	0.69	<b>0.82</b>

Table 45: Constituent ranking for the English-German Europarl Task

	CMU-SMT	CUED	CUED-C	LIMSI	RBMT3	RBMT4	RBMT5	RBMT6	SAAR	UCB	UEDIN	UPC
CMU-SMT		0.19	0.17	<b>0.26</b>	<b>0.38</b>	0.27	<b>0.45</b>	0.32	<b>0.35</b>	<b>0.27</b>	<b>0.26</b>	<b>0.2</b>
CUED	<b>0.21</b>		<b>0.21</b>	0.24	0.24	0.2	<b>0.34</b>	0.25	0.27	0.18	<b>0.26</b>	0.21
CUED-CONTRAST	0.17	0.08		0.12	0.24	0.23*	0.27	0.25	0.21	0.12	0.11	<b>0.26</b>
LIMSI	0.17	<b>0.25</b>	<b>0.26</b>		0.34	0.18 <sup>†</sup>	0.33	0.33	<b>0.31</b>	0.17	<b>0.26</b>	<b>0.23</b>
RBMT3	0.29	<b>0.31</b>	<b>0.35</b>	<b>0.37</b>		0.21	<b>0.4</b>	<b>0.31</b>	<b>0.32</b>	<b>0.43</b>	<b>0.42</b>	<b>0.52*</b>
RBMT4	<b>0.38</b>	<b>0.34</b>	<b>0.54*</b>	<b>0.47<sup>†</sup></b>	<b>0.35</b>		<b>0.24</b>	<b>0.32</b>	<b>0.46<sup>†</sup></b>	<b>0.37</b>	<b>0.40</b>	<b>0.53</b>
RBMT5	0.24	0.31	<b>0.40</b>	0.33	0.25	0.18		<b>0.31</b>	<b>0.33</b>	0.32	0.28	0.38
RBMT6	<b>0.33</b>	<b>0.29</b>	<b>0.28</b>	0.33	0.26	0.27	0.16		<b>0.26</b>	0.3	<b>0.39</b>	<b>0.41</b>
SAAR	0.26	0.27	<b>0.33</b>	0.26	0.21	0.12 <sup>†</sup>	0.25	0.24		0.20	<b>0.28</b>	0.20
UCB	0.25	<b>0.30</b>	<b>0.23</b>	<b>0.27</b>	0.31	0.27	<b>0.40</b>	<b>0.34</b>	<b>0.28</b>		<b>0.32</b>	0.26
UEDIN	0.19	0.20	<b>0.19</b>	0.24	0.27	0.33	<b>0.31</b>	0.27	0.21	0.21		<b>0.25</b>
UPC	0.1	0.21	0.17	0.2	0.22*	0.28	<b>0.4</b>	0.24	<b>0.29</b>	<b>0.30</b>	0.2	
> OTHERS	0.24	0.25	0.28	0.28	0.28	0.23	<b>0.33</b>	0.29	0.3	0.26	0.3	0.32
≥ OTHERS	0.72	0.76	<b>0.82</b>	0.74	0.64	0.61	0.7	0.70	0.76	0.71	0.76	0.76

Table 46: Constituent ranking for the Spanish-English News Task

	CMU-SMT	CUED	DCU	LIMSI	RBMT3	RBMT4	RBMT5	RBMT6	SAAR	UCL	UEDIN	UPC
CMU-SMT		<b>0.2</b>	<b>0.20</b>	<b>0.1</b>	0.1 <sup>†</sup>	0.18 <sup>†</sup>	0.04 <sup>†</sup>	0.18 <sup>†</sup>	<b>0.16</b>	0.17	<b>0.19</b>	0.19
CUED	0.18		0.13	<b>0.19</b>	0.14 <sup>†</sup>	0.12 <sup>†</sup>	0.1 <sup>†</sup>	0.2*	<b>0.13</b>	0.12*	<b>0.22</b>	0.12
DCU	0.15	0.13		0.11	0.09 <sup>†</sup>	0.10 <sup>†</sup>	0.13 <sup>†</sup>	0.09 <sup>†</sup>	<b>0.19</b>	0.15*	0.14	0.15
LIMSI	0.03	0.15	<b>0.16</b>		0.19 <sup>†</sup>	0.18 <sup>†</sup>	0.15 <sup>†</sup>	0.19 <sup>†</sup>	<b>0.19</b>	0.08 <sup>†</sup>	0.07	<b>0.22</b>
RBMT3	<b>0.7<sup>†</sup></b>	<b>0.73<sup>†</sup></b>	<b>0.59<sup>†</sup></b>	<b>0.49<sup>†</sup></b>		0.19	<b>0.36</b>	0.22	<b>0.62<sup>†</sup></b>	<b>0.55*</b>	<b>0.68<sup>†</sup></b>	<b>0.73<sup>†</sup></b>
RBMT4	<b>0.55<sup>†</sup></b>	<b>0.62<sup>†</sup></b>	<b>0.51<sup>†</sup></b>	<b>0.55<sup>†</sup></b>	<b>0.23</b>		0.22	0.17	<b>0.56<sup>†</sup></b>	<b>0.43</b>	<b>0.56<sup>†</sup></b>	<b>0.44*</b>
RBMT5	<b>0.60<sup>†</sup></b>	<b>0.61<sup>†</sup></b>	<b>0.53<sup>†</sup></b>	<b>0.61<sup>†</sup></b>	0.32	<b>0.38</b>		<b>0.28</b>	<b>0.63<sup>†</sup></b>	<b>0.53</b>	<b>0.7<sup>†</sup></b>	<b>0.59<sup>†</sup></b>
RBMT6	<b>0.52<sup>†</sup></b>	<b>0.48*</b>	<b>0.51<sup>†</sup></b>	<b>0.49<sup>†</sup></b>	<b>0.23</b>	<b>0.26</b>	0.19		<b>0.49<sup>†</sup></b>	<b>0.53*</b>	<b>0.52<sup>†</sup></b>	<b>0.50<sup>†</sup></b>
SAAR	0.14	0.10	0.12	0.15	0.10 <sup>†</sup>	0.12 <sup>†</sup>	0.05 <sup>†</sup>	0.07 <sup>†</sup>		0.14*	0.05	<b>0.18</b>
UCL	<b>0.38</b>	<b>0.37*</b>	<b>0.46*</b>	<b>0.45<sup>†</sup></b>	0.28*	0.32	0.29	0.24*	<b>0.38*</b>		<b>0.38*</b>	<b>0.36</b>
UEDIN	0.06	0.14	0.14	<b>0.18</b>	0.15 <sup>†</sup>	0.16 <sup>†</sup>	0.05 <sup>†</sup>	0.16 <sup>†</sup>	<b>0.15</b>	0.10*		<b>0.21</b>
UPC	0.19	0.12	<b>0.20</b>	0.12	0.07 <sup>†</sup>	0.17*	0.09 <sup>†</sup>	0.14 <sup>†</sup>	0.04	0.17	0.14	
> OTHERS	0.32	0.33	0.32	0.32	0.17	0.2	0.15	0.17	0.33	0.28	0.34	<b>0.35</b>
≥ OTHERS	0.85	0.85	0.87	0.85	0.46	0.56	0.47	0.57	<b>0.89</b>	0.65	0.87	0.87

Table 47: Constituent ranking for the Spanish-English Europarl Task

	CMU-SMT	LIMSI	RBMT3	RBMT4	RBMT5	RBMT6	SAAR	UCB	UEDIN	UPC
CMU-SMT		0.20	<b>0.36</b>	<b>0.37</b>	0.24 <sup>†</sup>	0.36	0.32	0.21	0.17	0.27
LIMSI	<b>0.23</b>		<b>0.4</b>	<b>0.46*</b>	0.33	<b>0.39</b>	0.31	0.23	0.17	0.18
RBMT3	0.33	0.35		0.22	0.19 <sup>†</sup>	0.3	0.31	<b>0.49</b>	0.34	0.22
RBMT4	0.30	0.25*	<b>0.25</b>		0.17*	0.17*	0.24	0.19 <sup>†</sup>	0.34	0.30
RBMT5	<b>0.53<sup>†</sup></b>	<b>0.42</b>	<b>0.50<sup>†</sup></b>	<b>0.41*</b>		<b>0.35</b>	<b>0.50*</b>	<b>0.44</b>	0.37	0.29
RBMT6	0.36	0.35	<b>0.34</b>	<b>0.39*</b>	0.32		0.35	0.36	0.37	0.38
SAAR	<b>0.33</b>	<b>0.36</b>	<b>0.38</b>	<b>0.28</b>	0.24*	<b>0.38</b>		0.29	0.22*	0.24
UCB	<b>0.32</b>	<b>0.29</b>	0.35	<b>0.54<sup>†</sup></b>	0.33	<b>0.45</b>	<b>0.31</b>		0.19	<b>0.29</b>
UEDIN	<b>0.29</b>	<b>0.33</b>	<b>0.36</b>	<b>0.42</b>	<b>0.42</b>	<b>0.39</b>	<b>0.45*</b>	<b>0.30</b>		<b>0.44</b>
UPC	<b>0.36</b>	<b>0.42</b>	<b>0.50</b>	<b>0.49</b>	<b>0.42</b>	<b>0.44</b>	<b>0.51</b>	0.21	0.26	
> OTHERS	0.34	0.33	0.38	<b>0.39</b>	0.29	0.35	0.36	0.31	0.27	0.29
≥ OTHERS	0.72	0.69	0.69	<b>0.75</b>	0.57	0.64	0.7	0.65	0.63	0.6

Table 48: Constituent ranking for the English-Spanish News Task

	CMU-SMT	LIMSI	RBMT3	RBMT4	RBMT5	RBMT6	SAAR	UCL	UEDIN	UPC	UW
CMU-SMT		0.13	0.10 <sup>†</sup>	0.21*	0.2 <sup>†</sup>	0.2 <sup>†</sup>	<b>0.26</b>	0.22	0.13	0.16	0.14
LIMSI	<b>0.17</b>		0.24	0.16 <sup>†</sup>	0.20 <sup>†</sup>	0.13 <sup>†</sup>	<b>0.21</b>	0.06 <sup>†</sup>	0.09	<b>0.14</b>	0.08
RBMT3	<b>0.64<sup>†</sup></b>	<b>0.45</b>		0.24	0.30	0.21	<b>0.57<sup>†</sup></b>	<b>0.56</b>	<b>0.58*</b>	0.32	<b>0.58<sup>†</sup></b>
RBMT4	<b>0.54*</b>	<b>0.52<sup>†</sup></b>	<b>0.42</b>		0.26	0.24	<b>0.50*</b>	0.35	<b>0.43</b>	<b>0.47</b>	<b>0.44</b>
RBMT5	<b>0.61<sup>†</sup></b>	<b>0.68<sup>†</sup></b>	<b>0.46</b>	<b>0.44</b>			<b>0.64<sup>†</sup></b>	<b>0.50</b>	<b>0.63<sup>†</sup></b>	<b>0.62<sup>†</sup></b>	<b>0.54</b>
RBMT6	<b>0.57<sup>†</sup></b>	<b>0.48<sup>†</sup></b>	<b>0.39</b>	<b>0.33</b>	0.25		<b>0.52<sup>†</sup></b>	0.33	<b>0.54<sup>†</sup></b>	<b>0.46</b>	<b>0.46</b>
SAAR	0.19	0.14	0.07 <sup>†</sup>	0.19*	0.09 <sup>†</sup>	0.14 <sup>†</sup>		0.13 <sup>†</sup>	0.17	0.26	0.18
UCL	<b>0.43</b>	<b>0.46<sup>†</sup></b>	0.29	<b>0.37</b>	0.38	<b>0.42</b>	<b>0.49<sup>†</sup></b>		<b>0.37*</b>	<b>0.48</b>	<b>0.40</b>
UEDIN	<b>0.15</b>	<b>0.11</b>	0.24*	0.20	0.13 <sup>†</sup>	0.17 <sup>†</sup>	<b>0.30</b>	0.14*		0.20	0.20
UPC	<b>0.26</b>	0.05	<b>0.35</b>	0.25	0.16 <sup>†</sup>	0.23	<b>0.34</b>	0.21	<b>0.23</b>		0.10
UW	0.14	<b>0.14</b>	0.17 <sup>†</sup>	0.22	0.23	0.2	<b>0.32</b>	0.20	0.20	<b>0.35</b>	
> OTHERS	0.37	0.32	0.28	0.26	0.22	0.23	<b>0.42</b>	0.27	0.35	0.35	0.33
≥ OTHERS	0.83	<b>0.86</b>	0.56	0.59	0.46	0.57	0.85	0.59	0.82	0.78	0.79

Table 49: Constituent ranking for the English-Spanish Europarl Task

	CU-BOJAR	CU-TECTOMT	PC-TRANSLATOR	UEDIN
CU-BOJAR		0.33	<b>0.41</b>	0.28*
CU-TECTOMT	<b>0.37</b>		<b>0.42<sup>†</sup></b>	0.36
PC-TRANSLATOR	0.34	0.31 <sup>†</sup>		0.32 <sup>†</sup>
UEDIN	<b>0.37*</b>	<b>0.37</b>	<b>0.43<sup>†</sup></b>	
> OTHERS	0.36	0.34	<b>0.42</b>	0.32
≥ OTHERS	0.66	0.62	<b>0.67</b>	0.61

Table 50: Constituent ranking for the English-Czech News Task

	CU-BOJAR	CU-TECTOMT	PC-TRANSLATOR	UEDIN
CU-BOJAR		0.25 <sup>†</sup>	0.33 <sup>†</sup>	0.22 <sup>†</sup>
CU-TECTOMT	<b>0.50<sup>†</sup></b>		<b>0.44<sup>†</sup></b>	<b>0.45</b>
PC-TRANSLATOR	<b>0.47<sup>†</sup></b>	0.3 <sup>†</sup>		<b>0.40</b>
UEDIN	<b>0.39<sup>†</sup></b>	0.37	0.39	
> OTHERS	<b>0.45</b>	0.31	0.39	0.36
≥ OTHERS	<b>0.73</b>	0.54	0.61	0.61

Table 51: Constituent ranking for the English-Czech Commentary Task

French–English						English–French					
Europarl	YES	NO	News	YES	NO	Europarl	YES	NO	News	YES	NO
CMU-XFR	0.61	0.39	CMU-XFR	0.55	0.45	LIMSI	0.75	0.26	LIMSI	0.73	0.27
CUED	0.83	0.17	CUED	0.74	0.26	LIUM-SYS	<b>0.84</b>	<b>0.16</b>	LIUM-SYS	<b>0.75</b>	<b>0.25</b>
DCU	0.88	0.12	CUED-C	0.79	0.21	RBMT3	0.49	0.51	RBMT3	0.59	0.41
LIMSI	0.89	0.11	LIMSI	<b>0.81</b>	<b>0.2</b>	RBMT4	0.50	0.5	RBMT4	0.59	0.41
LIUM-SYS	0.89	0.11	LIUM-SYS	0.79	0.21	RBMT5	0.44	0.56	RBMT5	0.64	0.36
RBMT3	0.54	0.47	LI-SYS-C	0.7	0.30	RBMT6	0.35	0.65	RBMT6	0.58	0.42
RBMT4	0.62	0.38	RBMT3	0.63	0.37	SAAR	0.70	0.3	SAAR	0.59	0.41
RBMT5	0.71	0.29	RBMT4	0.64	0.36	UCL	0.6	0.40	SAAR-C	0.59	0.41
RBMT6	0.54	0.46	RBMT5	0.76	0.24	UEDIN	0.75	0.25	UEDIN	0.63	0.37
SAAR	0.72	0.28	RBMT6	0.66	0.34				XEROX	0.30	0.7
SAAR-C	0.86	0.14	RBMT5	0.64	0.36						
SYSTRAN	0.81	0.19	SAAR	0.64	0.36						
UCL	0.73	0.27	SAAR-C	0.70	0.3						
UEDIN	<b>0.91</b>	<b>0.09</b>	UEDIN	0.72	0.28						

German–English						English–German					
Europarl	YES	NO	News	YES	NO	Europarl	YES	NO	News	YES	NO
CMU-XFER	0.53	0.47	CMU-XFER	0.47	0.53	CMU-GIMPEL	0.82 <sup>†</sup>	0.18	LIMSI	0.56	0.44
LIMSI	0.80	0.2	LIMSI	0.73	0.28	LIMSI	0.79 <sup>†</sup>	0.21	LIU	0.49	0.51
LIU	0.83	0.17	LIU	0.64	0.36	LIU	0.79 <sup>†</sup>	0.21	RBMT2	0.69	0.31
RBMT2	0.76	0.24	RBMT2	0.72	0.28	RBMT2	0.69 <sup>†</sup>	0.31	RBMT3	0.69	0.31
RBMT3	0.74	0.26	RBMT3	0.73	0.27	RBMT3	0.57	0.43	RBMT4	<b>0.75</b>	<b>0.25</b>
RBMT4	0.67	0.33	RBMT4	<b>0.74</b>	<b>0.26</b>	RBMT4	0.67 <sup>†</sup>	0.34	RBMT5	0.55	0.45
RBMT5	0.63	0.37	RBMT5	0.59	0.41	RBMT5	0.45	0.55	RBMT6	0.6	0.40
RBMT6	0.63	0.37	RBMT6	0.68	0.32	RBMT6	0.47	0.53	SAAR	0.54	0.46
SAAR	0.82	0.18	SAAR	0.67	0.33	SAAR	0.77 <sup>†</sup>	0.23	UEDIN	0.52	0.48
UCL	0.49	0.51	SAAR-C	0.72	0.28	UCL	0.61 <sup>†</sup>	0.39			
UEDIN	<b>0.86</b>	<b>0.14</b>	UEDIN	0.63	0.37	UEDIN	<b>0.85<sup>†</sup></b>	<b>0.15</b>			

Spanish–English						English–Spanish					
Europarl	YES	NO	News	YES	NO	Europarl	YES	NO	News	YES	NO
CMU-SMT	0.88	0.12	CMU-SMT	0.64	0.37	CMU-SMT	0.80	0.2	CMU-SMT	0.46	0.54
CUED	0.86	0.14	CUED	0.64	0.36	LIMSI	0.87	0.13	LIMSI	0.53	0.47
DCU	0.85	0.15	CUED-C	0.69	0.31	RBMT3	0.58	0.42	RBMT3	0.64	0.36
LIMSI	<b>0.90</b>	<b>0.1</b>	LIMSI	0.68	0.33	RBMT4	0.6	0.40	RBMT4	<b>0.76</b>	<b>0.24</b>
RBMT3	0.65	0.35	RBMT3	0.61	0.39	RBMT5	0.64	0.37	RBMT5	0.6	0.40
RBMT4	0.56	0.44	RBMT4	0.65	0.35	RBMT6	0.60	0.40	RBMT6	0.62	0.38
RBMT5	0.59	0.41	RBMT5	0.59	0.41	SAAR	0.81	0.19	SAAR	0.64	0.36
RBMT6	0.55	0.45	RBMT6	0.64	0.37	UCL	0.71	0.29	UCB	0.57	0.43
SAAR	0.87	0.13	SAAR	0.7	0.30	UEDIN	0.89	0.11	UEDIN	0.49	0.51
UCL	0.73	0.27	UCB	0.64	0.37	UPC	<b>0.90</b>	<b>0.1</b>	UPC	0.37	0.63
UEDIN	0.88	0.12	UEDIN	0.62	0.38	UW	0.79	0.22			
UPC	0.86	0.14	UPC	<b>0.71</b>	<b>0.29</b>						

English–Czech					
Commentary	YES	NO	News	YES	NO
CU-BOJAR	<b>0.59</b>	<b>0.41</b>	CU-BOJAR	<b>0.54</b>	<b>0.46</b>
CU-TECTO	0.43	0.57	CU-TECTO	0.42	0.58
PC-TRANS	0.51	0.49	PC-TRANS	0.52	0.48
UEDIN	0.41	0.59	UEDIN	0.44	0.56

Table 52: Yes/No Acceptability of Constituents

