The TUNA-REG Challenge 2009: Overview and Evaluation Results

Albert Gatt
Computing Science
University of Aberdeen
Aberdeen AB24 3UE, UK
a.gatt@abdn.ac.uk

Anja Belz **Eric Kow**
Natural Language Technology Group
University of Brighton
Brighton BN2 4GJ, UK
{asb,eykk10}@bton.ac.uk

Abstract

The TUNA-REG'09 Challenge was one of the shared-task evaluation competitions at Generation Challenges 2009. TUNA-REG'09 used data from the TUNA Corpus of paired representations of entities and human-authored referring expressions. The shared task was to create systems that generate referring expressions for entities given representations of sets of entities and their properties. Four teams submitted six systems to TUNA-REG'09. We evaluated the six systems and two sets of human-authored referring expressions using several automatic intrinsic measures, a human-assessed intrinsic evaluation and a human task performance experiment. This report describes the TUNA-REG task and the evaluation methods used, and presents the evaluation results.

1 Introduction

This year's run of the TUNA-REG Shared-Task Evaluation Competition (STEC) is the third, and final, competition to involve the TUNA Corpus of referring expressions. The TUNA Corpus was first used in the Pilot Attribute Selection for Generating Referring Expressions (ASGRE) Challenge (Belz and Gatt, 2007) which took place between May and September 2007; and again for three of the shared tasks in Referring Expression Generation (REG) Challenges 2008, which ran between September 2007 and May 2008 (Gatt et al., 2008). This year's TUNA Task replicates one of the three tasks from REG'08, the TUNA-REG Task. It uses the same test data, to enable direct comparison against the 2008 results. Four participating teams submitted 6 different systems this year; teams and their affiliations are shown in Table 1.

Team ID	Affiliation
GRAPH	Macquarie, Tilburg and Twente Universities
IS	ICSI, University of California
NIL-UCM	Universidad Complutense de Madrid
USP	University of São Paolo

Table 1: TUNA-REG'09 Participants.

2 Data

Each file in the TUNA corpus[1] consists of a single pairing of a domain (a representation of 7 entities and their attributes) and a human-authored description for one of the entities (the target referent). Some domains represent sets of people, some represent items of furniture (see also Table 2). The descriptions were collected in an online elicitation experiment which was advertised mainly on a website hosted at the University of Zurich Web Experimentation List[2] (a web service for recruiting subjects for experiments), and in which participation was not controlled or monitored. In the experiment, participants were shown pictures of the entities in the given domain and were asked to type a description of the target referent (which was highlighted in the visual display). The main condition[3] manipulated in the experiment was $+/-LOC$: in the $+LOC$ condition, participants were told that they could refer to entities using any of their properties (including their location on the screen). In the $-LOC$ condition, they were discouraged from doing so, though not prevented.

The XML format we have been using in the TUNA-REG STECs, shown in Figure 1, is a variant of the original format of the TUNA corpus. The root TRIAL node has a unique ID and an indication of the $+/-LOC$ experimental condi-

[1] http://www.csd.abdn.ac.uk/research/tuna/
[2] http://genpsylab-wexlist.unizh.ch
[3] The elicitation experiment had an additional independent variable, manipulating whether descriptions were elicited in a 'fault-critical' or 'non-fault-critical' condition. For the shared tasks this was ignored by collapsing all the data in these two conditions.

Proceedings of the 12th European Workshop on Natural Language Generation, pages 174–182,
Athens, Greece, 30 – 31 March 2009.

tion. The DOMAIN node contains 7 ENTITY nodes, which themselves contain a number of ATTRIBUTE nodes defining the possible properties of an entity in attribute-value notation. The attributes include properties such as an object's colour or a person's clothing, and the location of the image in the visual display which the DOMAIN represents. Each ENTITY node indicates whether it is the target referent or one of the six distractors, and also has a pointer to the image that it represents. The WORD-STRING is the actual description typed by one of the human authors, the ANNOTATED-WORD-STRING is the description with substrings annotated with the attributes they realise, while the ATTRIBUTE-SET contains the set of attributes only. The ANNOTATED-WORD-STRING and ATTRIBUTE-SET nodes were provided in the training and development data only, to show how substrings of a human-authored description mapped to attributes.

```
<TRIAL CONDITION="+/-LOC" ID="...">
  <DOMAIN>
    <ENTITY ID="..." TYPE="target" IMAGE="...">
      <ATTRIBUTE NAME="..." VALUE="..." />
      ...
    </ENTITY>
    <ENTITY ID="..." TYPE="distractor" IMAGE="...">
      <ATTRIBUTE NAME="..." VALUE="..." />
      ...
    </ENTITY>
    ...
  </DOMAIN>
  <WORD-STRING>
    string describing the target referent
  </WORD-STRING>
  <ANNOTATED-WORD-STRING>
    string in WORD-STRING annotated
    with attributes in ATTRIBUTE-SET
  </ANNOTATED-WORD-STRING>
  <ATTRIBUTE-SET>
    set of domain attributes in the description
  </ATTRIBUTE-SET>
</TRIAL>
```

Figure 1: XML format of corpus items.

Apart from differences in the XML format, the data used in the TUNA-REG Task also differs from the original TUNA corpus in that it has only the singular referring expressions from the original corpus, and in that we have added to it the files of images of entities that the XML mark-up points to.

The test set, which was constructed for the 2008 run of the TUNA-REG Task, consists of 112 items, each with a different domain paired with *two* human-authored descriptions. The items are distributed equally between furniture items and people, and between both experimental conditions ($+/-LOC$). In the following sections, the two sets of human descriptions will be referred to as HUMAN-1 and HUMAN-2.[4] The numbers of files in the training, development and test sets, as well as in the people and furniture subdomains, are shown in Table 2.

	Furniture	**People**	**All**
Training	319	274	593
Development	80	68	148
Test	56	56	112
All	455	398	853

Table 2: TUNA-REG data: subset sizes.

3 The TUNA-REG Task

Referring Expression Generation (REG) has been the subject of intensive research in the NLG community, giving rise to substantial consensus on the problem definition, as well as the nature of the inputs and outputs of REG algorithms. Typically, such algorithms take as input a domain, consisting of entities and their attributes, together with an indication of which is the intended referent, and output a set of attributes true of the referent which distinguish it from other entities in the domain. The TUNA-REG task adds an additional stage (realisation) in which selected attributes are mapped to a natural language expression (usually a noun phrase). Realisation has received far less attention among REG researchers than attribute selection.

The TUNA-REG task is an 'end-to-end' referring expression generation task, in the sense that it takes as input a representation of a set of entities and their properties, and outputs a word string which describes the target entity. Participating systems were not constrained to have attribute selection as a separate module from realisation.

In terms of the XML format, the items in the test set distributed to participants consisted of a DOMAIN node and ATTRIBUTE-SET, and participating systems had to generate appropriate WORD-STRINGs.

As with previous STECs involving the TUNA data, we deliberately refrained from including in the task definition any aim that would imply assumptions about quality (as would be the case if we had asked participants to aim to produce, say, minimal or uniquely distinguishing referring expressions), and instead we simply listed the evaluation criteria that were going to be used (described in Section 5).

[4] Descriptions in each set are not all by the same author.

Evaluation criterion	*Type of evaluation*	*Evaluation technique*
Humanlikeness	Intrinsic/automatic	Accuracy, String-edit distance, BLEU-3, NIST
Adequacy/clarity	Intrinsic/human	Judgment of adequacy as rated by native speakers
Fluency	Intrinsic/human	Judgment of fluency as rated by native speakers
Referential clarity	Extrinsic/human	Speed and accuracy in identification experiment

Table 3: Overview of evaluation methods.

4 Participating Teams and Systems

This section briefly describes this year's submissions. Full descriptions of participating systems can be found in the participants' reports included in this volume.

IS: The submission of the IS team, IS-FP-GT, is based on the idea that different writers use different styles of referring expressions, and that, therefore, knowing the identity of the writer helps generate REs similar to those in the corpus. The attribute-selection algorithm is an extended full-brevity algorithm which uses a nearest neighbour technique to select the attribute set (AS) most similar to a given writer's previous ASs, or, in a case where no ASs by the given writer have previously been seen, to select the AS that has the highest degree of similarity with all previously seen ASs by any writer. If multiple ASs remain, the algorithm first selects the shortest, then the most representative of the remaining REs, then the AS with the highest-frequency attributes. Individualised statistical models are used to convert the selected AS into a surface-syntactic dependency tree which is then converted to a word stirng with an existing realiser.

GRAPH: The GRAPH team reused their existing graph-based attribute selection component, which represents a domain as a weighted graph, and uses a cost function for attributes. The team developed a new realiser which uses a set of templates derived from the descriptions in the TUNA corpus. In order to build templates, certain subsets of attributes were grouped together, individual attributes were replaced by their type, and a preferred order for attributes was determined based on frequencies of orderings. During realisation, if a matching template exists, types are replaced with the most frequent word string for each given attribute; if no match exists, realisation is done by a simple rule-based method.

NIL-UCM: The three systems submitted by this group use a standard evolutionary algorithm for attribute selection where genotypes consist of binary-valued genes each representing the presence or absence of a given attribute. Realisation is done with a case-based reasoning (CBR) method which retrieves the most similar previously seen ASs for an input AS, in order of their similarity to the input. (Sub)strings are then copied from the preferred retrieved case to create the output word string. One system, NIL-UCM-EvoCBR uses both components as described above. The other two systems, NIL-UCM-ValuesCBR and NIL-UCM-EvoTAP, replace one of the components with the team's corresponding component from REG'08.

USP: The system submitted by this group, USP-EACH, is a frequency-based greedy attribute selection strategy which takes into account the $+/-$ *LOC* attribute in the TUNA data. Realisation was done using the surface realiser supplied to participants in the ASGRE'07 Challenge.

5 Evaluation Methods and Results

We used a range of different evaluation methods, including intrinsic and extrinsic,[5] automatically computed and human-evaluated, as shown in the overview in Table 3. Participants computed automatic intrinsic evaluation scores on the development set (using the `teval` program provided by us). We performed all of the evaluations shown in Table 3 on the test data set. For all measures, results were computed both (a) overall, using the entire test data set, and (b) by entity type, that is, computing separate values for outputs in the *furniture* and in the *people* domain. Evaluation methods for each evaluation type and corresponding evaluation results are presented in the following three sections.

5.1 Automatic intrinsic evaluations

Humanlikeness, by which we mean the similarity of system outputs to sets of human-produced reference 'outputs', was assessed using Accuracy,

[5] Intrinsic evaluations assess properties of peer systems in their own right, whereas extrinsic evaluations assess the effect of a peer system on something that is external to it, such as its effect on human performance at some given task or the added value it brings to an application.

	All development data			People			Furniture		
	Accuracy	SE	BLEU	Accuracy	SE	BLEU	Accuracy	SE	BLEU
IS-FP-GT	9.71%	4.313	0.297	4.41%	4.764	0.2263	15%	3.863	0.3684
GRAPH	–	5.03	0.30	–	5.15	0.33	–	4.94	0.27
NIL-UCM-EVOTAP	6%	5.41	0.20	3%	6.04	0.15	8%	4.87	0.24
NIL-UCM-VALUESCBR	1%	5.86	0.19	1%	5.80	0.17	1%	5.91	0.20
USP-EACH	–	6.03	0.19	–	7.50	0.04	–	4.78	0.31
NIL-UCM-EVOCBR	3%	6.31	0.17	1%	6.94	0.16	4%	5.77	0.18

Table 4: Participating teams' self-reported automatic intrinsic scores on development data set with single human-authored reference description (listed in order of overall mean SE score).

	All test data				People				Furniture			
	Acc	SE	BLEU	NIST	Acc	SE	BLEU	NIST	Acc	SE	BLEU	NIST
GRAPH	12.50	6.41	0.47	2.57	8.93	7.04	0.43	2.16	16.07	5.79	0.51	2.26
IS-FP-GT	3.57	6.74	0.28	0.75	3.57	7.04	0.37	0.94	3.57	6.45	0.13	0.36
NIL-UCM-EVOTAP	6.25	7.28	0.26	0.90	3.57	8.07	0.20	0.45	8.93	6.48	0.34	1.22
USP-EACH	7.14	7.59	0.27	1.33	0.00	9.04	0.11	0.46	14.29	6.14	0.41	2.28
NIL-UCM-VALUESCBR	2.68	7.71	0.27	1.69	3.57	8.07	0.23	0.94	1.79	7.34	0.28	1.99
NIL-UCM-EVOCBR	2.68	8.02	0.26	1.97	0.00	9.07	0.19	1.65	5.36	6.96	0.35	1.69
HUMAN-2	2.68	9.68	0.12	1.78	3.57	10.64	0.12	1.50	1.79	8.71	0.13	1.57
HUMAN-1	2.68	9.68	0.12	1.68	3.57	10.64	0.12	1.41	1.79	8.71	0.12	1.49

Table 5: Automatic intrinsic scores on test data set with two human-authored reference descriptions (listed in order of overall mean SE score).

string-edit distance, BLEU-3 and NIST-5. Accuracy measures the percentage of cases where a system's output word string was identical to the corresponding description in the corpus. String-edit distance (SE) is the classic Levenshtein distance measure and computes the minimal number of insertions, deletions and substitutions required to transform one string into another. We set the cost for insertions and deletions to 1, and that for substitutions to 2. If two strings are identical, then this metric returns 0 (perfect match). Otherwise the value depends on the length of the two strings (the maximum value is the sum of the lengths). As an aggregate measure, we compute the mean of pairwise SE scores.

BLEU-x is an n-gram based string comparison measure, originally proposed by Papineni et al. (2001; 2002) for evaluation of Machine Translation systems. It computes the proportion of word n-grams of length x and less that a system output shares with several reference outputs. Setting $x = 4$ (i.e. considering all n-grams of length ≤ 4) is standard, but because many of the TUNA descriptions are shorter than 4 tokens, we compute BLEU-3 instead. BLEU ranges from 0 to 1.

NIST is a version of BLEU, but where BLEU gives equal weight to all n-grams, NIST gives more importance to less frequent n-grams, which are taken to be more informative. The maximum NIST score depends on the size of the test set.

Unlike string-edit distance, BLEU and NIST are by definition aggregate measures (i.e. a single score is obtained for a peer system based on the entire set of items to be compared, and this is not generally equal to the average of scores for individual items).

Because the test data has two human-authored reference descriptions per domain, the Accuracy and SE scores had to be computed slightly differently to obtain test data scores (whereas BLEU and NIST are designed for multiple reference texts). For the test data only, therefore, Accuracy expresses the percentage of a system's outputs that match at least *one* of the reference outputs, and SE is the average of the two pairwise scores against the reference outputs.

Results: Table 4 is an overview of the self-reported scores on the development set included in the participants' reports (not all participants report Accuracy scores). The corresponding scores for the test data set as well as NIST scores for the test data (all computed by us), are shown in Table 5. The table also includes the result of comparing the two sets of human descriptions, HUMAN-1 and HUMAN-2, to each other using the same metrics (their scores are distinct only for non-commutative measures, i.e. NIST and BLEU).

We ran[6] a one-way ANOVA for the SE scores.

[6]We used SPSS for all statistical analyses and tests.

There was a main effect of SYSTEM on SE ($F = 10.938, p < .001$). A post-hoc Tukey HSD test with $\alpha = .05$ revealed a number of significant differences: all systems were significantly better than the human-authored descriptions, and GRAPH was furthermore significantly better than NIL-UCM-EVOCBR.

We also computed the Kruskal-Wallis H value for the systems' individual Accuracy scores, using a chi square test to establish significance. By this test, the observed aggregate difference among the seven systems is significant at the .01 level ($\chi^2_7 = 20.169$).

5.2 Human intrinsic evaluation

The TUNA'09 Challenge was the first TUNA shared-task competition to include an intrinsic evaluation involving human judgments of quality.

Design: The intrinsic human evaluation involved descriptions for all 112 test data items from all six submitted systems, as well as from the two sets of human-authored descriptions.[7] Thus, each of the 112 test set items was associated with 8 different descriptions. We used a Repeated Latin Squares design which ensures that each subject sees descriptions from each system and for each domain the same number of times. There were fourteen 8×8 squares, and a total of 896 individual judgments in this evaluation, each system receiving 112 judgments (14 from each subject).

Procedure: In each of the 112 trials, participants were shown a system output (i.e. a `WORD-STRING`), together with its corresponding domain, displayed as the set of corresponding images on the screen.[8] The intended (target) referent was highlighted by a red frame surrounding it on the screen. They were asked to give two ratings in answer to the following questions (the first for *Adequacy*, the second for *Fluency*):

1. *How clear is this description?* Try to imagine someone who could see the same grid with the same pictures, but didn't know which of the pictures was the target. How easily would they be able to find it, based on the phrase given?
2. *How fluent is this description?* Here your task is to judge how well the phrase reads. Is it good, clear English?

We did not use a rating scale (where integers correspond to different assessments of quality), because it is not generally considered appropriate to apply parametric methods of analysis to ordinal data. Instead, we asked subjects to give their judgments for Adequacy and Fluency for each item by manipulating a slider like this:

The slider pointer was placed in the center at the beginning of each trial, as shown above. The position of the slider selected by the subject mapped to an integer value between 1 and 100. However, the scale was not visible to participants, whose task was to move the pointer to the left or right. The further to the right, the more positive the judgment (and the higher the value returned); the further to the left, the more negative.

Following instructions, subjects did two practice examples, followed by the 112 test items in random order. Subjects carried out the experiment over the internet, at a time and place of their choosing, and were allowed to interrupt and resume the experiment. According to self-reported timings, subjects took between 25 and 60 minutes to complete the experiment (not counting breaks).

Participants: We recruited eight native speakers of English from among post-graduate students currently doing a Masters degree in a linguistics-related subject.[9]

We recorded subjects' gender, level of education, field of study, proficiency in English, variety of English and colour vision. Since all subjects were native English speakers, had normal colour vision, and had comparable levels of education and academic backgrounds, as indicated above, these variables are not included in the analyses reported below.

Results: Table 6 displays the mean Fluency and Adequacy judgments obtained by each system. We conducted two separate 8 (SYSTEM) $\times$ 2 (DOMAIN) Univariate Analyses of Variance (ANOVAs) on Adequacy and Fluency, where DOMAIN ranges

[7] Note that we refer to all outputs, whether human or system-generated, as *system outputs* in what follows.

[8] The on-screen display of images was very similar, although not identical, to that in the original TUNA elicitation experiments.

[9] MA Linguistics and MRes Speech, Language and Cognition at UCL; MA Applied Linguistics and MRes Psychology at Sussex; and MA Media-assisted Language Teaching at Brighton.

	All test data				People				Furniture			
	Adequacy		Fluency		Adequacy		Fluency		Adequacy		Fluency	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD
GRAPH	84.11	21.07	85.81	17.52	85.30	18.10	87.70	14.42	82.91	23.78	83.93	20.11
USP-EACH	77.72	28.33	84.20	20.27	81.04	26.48	81.82	24.47	74.41	29.93	86.57	14.79
NIL-UCM-EVOTAP	76.16	28.34	61.95	26.13	78.66	27.48	59.13	29.78	73.66	29.22	64.77	21.79
HUMAN-2	74.63	34.77	73.38	27.63	80.93	31.83	73.16	30.88	68.34	36.68	73.59	24.23
NIL-UCM-VALUESCBR	72.34	33.93	59.41	33.94	68.18	37.37	46.23	34.92	76.50	29.86	72.59	27.43
HUMAN-1	70.38	34.92	71.52	30.79	83.39	24.27	72.39	28.55	57.36	39.08	70.64	33.13
NIL-UCM-EVOCBR	63.65	37.19	55.38	35.32	56.61	40.20	41.45	37.38	70.70	32.76	69.30	26.93
IS-FP-GT	59.46	40.94	66.21	30.97	88.79	19.26	65.27	32.22	30.14	35.51	67.16	29.94

Table 6: Human-assessed intrinsic scores on test data set, including the two sets of human-authored reference descriptions (listed in order of overall mean Adequacy score).

Adequacy					Fluency					
GRAPH	A				GRAPH	A				
USP-EACH	A	B			USP-EACH	A	B			
NIL-UCM-EVOTAP	A	B			HUMAN-2		B	C		
HUMAN-2	A	B	C		HUMAN-1			C	D	
NIL-UCM-VALUESCBR	A	B	C		IS-FP-GT			C	D	E
HUMAN-1		B	C	D	NIL-UCM-EVOTAP				D	E
NIL-UCM-EVOCBR			C	D	NIL-UCM-VALUESCBR					E
IS-FP-GT				D	NIL-UCM-EVOCBR					E

Table 7: Homogeneous subsets for Adequacy and Fluency. Systems which do not share a letter are significantly different at $\alpha = .05$.

over People and Furniture Items. On Adequacy, there were main effects of SYSTEM ($F(7, 880) = 7.291, p < .001$) and DOMAIN ($F(1, 880) = 29.133, p < .001$), with a significant interaction between the two ($F(7, 880) = 15.30, p < .001$). On Fluency, there were main effects of SYSTEM ($F(7, 880) = 18.14$) and of DOMAIN ($F(7, 880) = 17.20$), again with a significant SYSTEM $\times$ DOMAIN interaction ($F(7, 880) = 5.60$), all significant at $p < .001$. Post-hoc Tukey comparisons on both dependent measures yielded the homogeneous subsets displayed in Table 7.

5.3 Extrinsic task-performance evaluation

As for earlier shared tasks involving the TUNA data, we carried out a task-performance experiment in which subjects have the task of identifying intended referents.

Design: The extrinsic human evaluation involved descriptions for all 112 test data items from all six submitted systems, as well as from the two sets of human-authored descriptions. We used a Repeated Latin Squares design with fourteen 8×8 squares, so again there were a total of 896 individual judgments and each system received 112 judgments, however this time it was 7 from each subject, as there were 16 participants; so half the participants did the first 56 items (the first 7 squares), and the other half the second 56 (the remaining 7 squares).

Procedure: In each of their 5 practice trials and 56 real trials, participants were shown a system output (i.e. a `WORD-STRING`), together with its corresponding domain, displayed as the set of corresponding images on the screen. In this experiment the intended referent was not highlighted in the on-screen display, and the participants' task was to identify the intended referent among the pictures by mouse-clicking on it.[10]

In previous TUNA identification experiments (Belz and Gatt, 2007; Gatt et al., 2008), subjects had to read the description before identifying the intended referent. In ASGRE'07 both description and pictures were displayed at the same time, yielding a single time measure that combined reading and identification times. In REG'08, subjects first read the description and then called up the pictures on the screen when they had finished reading the description, which yielded separate reading and identification times.

[10] Due to limitations related to the stimulus presentation software, the images in this experiment were displayed in strict rows and columns, whereas the display grid in the web-based TUNA elicitation experiment and the intrinsic human evalution experiment were slightly distorted. This may have affected timings in those (very rare) cases where a description explicitly referenced the column a target referent was located in, as in *the chair in column 1*.

This year we tried out a version of the experiment where subjects listened to descriptions read out by a synthetic voice[11] over headphones while looking at the pictures displayed on the screen.

Stimulus presentation was carried out using DMDX, a Win-32 software package for psycholinguistic experiments involving time measurements (Forster and Forster, 2003). Participants initiated each trial, which consisted of an initial warning bell and a fixation point flashed on the screen for 1000ms. Following this, the visual domain was displayed, and the voice reading the description was initiated after a delay of 500ms. We recorded time in milliseconds from the start of display to the mouse-click whereby a participant identified the target referent. This is hereafter referred to as the *identification speed*. The analysis reported below also uses *identification accuracy*, the percentage of correctly identified target referents, as an additional dependent variable. Trials timed out after $15,000$ms.

Participants: The experiment was carried out by 16 participants recruited from among the faculty and administrative staff of the University of Brighton. All participants carried out the experiment under supervision in the same quiet room on the same laptop, in the same ambient conditions, with no interruptions. All participants were native speakers, and we recorded type of post, whether they had normal colour vision and hearing, and whether they were left or right-handed.

Timeouts and outliers: None of the trials reached time-out stage during the experiment. Outliers were defined as those identification times which fell outside the *mean* ± 2SD (standard deviation) range. 44 data points (4.9%) out of a total of 896 were identified as outliers by this definition; these were replaced with the series mean (Ratliff, 1993). The results reported for identification speed below are based on these adjusted times.

Results: Table 8 displays mean identification speed and identification accuracy per system. A univariate ANOVA on identification speed revealed significant main effects of SYSTEM ($F(7,880) = 4.04, p < .001$) and DOMAIN ($F(1,880) = 11.53, p < .001$), with a significant interaction ($F(7,880) = 6.02, p < .001$). Table 9 displays homogeneous subsets obtained following pairwise comparisons using a post-hoc Tukey HSD analysis.

[11] We used the University of Edinburgh's Festival speech generation system (Black et al., 1999) in combination with the nitech_us_slt_arctic_hts voice, a high-quality female American voice.

USP-EACH	A	
GRAPH	A	
NIL-UCM-EVOTAP	A	B
IS-FP-GT	A	B
NIL-UCM-VALUESCBR	A	B
NIL-UCM-EVOCBR	A	B
HUMAN-2		B
HUMAN-1		B

Table 9: Homogeneous subsets for Identification Speed. Systems which do not share a letter are significantly different at $\alpha = .05$.

We treated identification accuracy as an indicator variable (indicating whether a participant correctly identified a target referent or not in a given trial). A Kruskal-Wallis test showed a significant difference between systems ($\chi^2_7 = 44.98; p < .001$).

5.4 Correlations

Table 10 displays the correlations between the eight evaluation measures we used. The numbers are Pearson product-moment correlation coefficients, calculated on the means (1 mean per system on each measure).

As regards the human-assessed intrinsic scores, there is no significant correlation between Adequacy and Fluency. Among the automatically computed intrinsic measures, the only significant correlation is between Accuracy and BLEU. For the extrinsic identification performance measures, there is no significant correlation between Identification Accuracy and Identification Speed.

As for correlations across the two types (human-assessed and automatically computed) of intrinsic measures, the only significant correlations are between Fluency and Accuracy, and between Adequacy and Accuracy. So, a system with a higher percentage of human-like outputs (as measured by Accurach) also tends to be scored more highly in terms of Fluency and Adequacy by humans.

We also found significant correlations between intrinsic and extrinsic measures: there was a strong and significant correlation between Identification Accuracy and Adequacy, implying that more adequate system outputs allowed people to identify target referents more correctly; there was also a significant (negative) correlation between

	All test data			People			Furniture		
	ID acc.	ID. speed		ID acc.	ID. speed		ID acc.	ID. speed	
	%	Mean	SD	%	Mean	SD	%	Mean	SD
GRAPH	0.96	3069.16	878.89	0.95	3081.01	767.62	0.96	3057.31	984.60
HUMAN-1	0.91	3517.58	1028.83	0.95	3323.76	764.59	0.88	3711.41	1214.55
USP-EACH	0.90	3067.16	821.00	0.86	3262.79	865.61	0.95	2871.53	730.15
NIL-UCM-EVOTAP	0.88	3159.41	910.65	0.88	3375.17	948.46	0.89	2943.65	824.17
NIL-UCM-VALUESCBR	0.87	3262.53	974.55	0.80	3447.50	1003.21	0.93	3077.56	916.87
HUMAN-2	0.83	3463.88	1001.29	0.89	3647.41	1045.95	0.77	3280.35	927.79
NIL-UCM-EVOCBR	0.81	3362.22	892.45	0.75	3779.64	831.91	0.88	2944.80	748.69
IS-FP-GT	0.68	3167.11	964.45	0.89	2980.30	750.78	0.46	3353.91	1114.68

Table 8: Identification speed and accuracy per system. Systems are displayed in descending order of overall identification accuracy.

	Human-assessed, intrinsic		Extrinsic		Auto-assessed, intrinsic			
	Fluency	Adequacy	ID Acc.	ID Speed	Acc.	SE	BLEU	NIST
Fluency	1	0.68	0.50	-0.89*	.85*	-0.57	0.66	0.30
Adequacy	0.68	1	0.95**	-0.65	.83*	-0.29	0.60	0.48
Identification Accuracy	0.50	0.95**	1	-0.39	0.68	-0.01	0.49	0.60
Identification Speed	0.89*	-0.65	-0.39	1	-0.79	0.68	-0.51	0.06
Accuracy	0.85*	0.83*	0.68	-0.79	1.00	-0.68	.859*	0.49
SE	-0.57	-0.29	-0.01	0.68	-0.68	1	-0.75	-0.07
BLEU	0.66	0.60	0.49	-0.51	.86*	-0.75	1	0.71
NIST	0.30	0.48	0.60	0.06	0.49	-0.07	0.71	1

Table 10: Correlations (Pearson's r) between all evaluation measures. (*significant at $p \leq .05$; **significant at $p \leq .01$)

Fluency and Identification Speed, implying that more fluent descriptions led to faster identification. While these results differ from previous findings (Belz and Gatt, 2008), in which no significant correlations were found between extrinsic measures and automatic intrinsic metrics, it is worth noting that significance in the results reported here was only observed between *human-assessed* intrinsic measures and the extrinsic ones.

6 Concluding Remarks

The three editions of the TUNA STEC have attracted a substantial amount of interest. In addition to a sizeable body of new work on referring expression generation, as another tangible outcome of these STECs we now have a wide range of different sets of system outputs for the same set of inputs. A particularly valuable resource is the pairing of these outputs from the submitted systems in each edition with evaluation data.

As this was the last time we are running a STEC with the TUNA data, we will now make all data sets, documentation and evaluation software from all TUNA STECs available to researchers. We are planning to add to these as many system outputs as we can, so that other researchers can perform evaluations involving these.

We are also planning to complete our evaluations of the evaluation methods we have developed. Among such experiments will be direct comparisons between the results of the three variants of the identification experiment we have tried out, and a direct comparison between different designs for human-assessed intrinsic evaluations (e.g. comparing the slider design reported here to preference judgments and rating scales).

Apart from the technological progress in REG which we hope the TUNA STECs have helped achieve, perhaps the single most important scientific result is strong evidence for the importance of extrinsic evaluations, as these do not necessarily agree with the results of much more commonly used intrinsic types of evaluations.

Acknowledgments

We thank our colleagues at the University of Brighton who participated in the identification experiment, and the Masters students at UCL, Sussex and Brighton who participated in the quality assessment experiment. The evaluations were funded by EPSRC (UK) grant EP/G03995X/1.

References

A. Belz and A. Gatt. 2007. The attribute selection for gre challenge: Overview and evaluation results. In

Proceedings of UCNLG+MT: Language Generation and Machine Translation.

A. Belz and A. Gatt. 2008. Intrinsic vs. extrinsic evaluation measures for referring expression generation. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics (ACL'08)*, pages 197–200.

A. Black, P. Taylor, and R. Caley, 1999. *The Festival Speech Synthesis System: System Documentation*. University of Edinburgh, 1.4 edition.

K. I. Forster and J. C. Forster. 2003. DMDX: A windows display program with millisecond accuracy. *Behavior Research Methods, Instruments, & Computers*, 35(1):116–124.

A. Gatt, A. Belz, and Eric Kow. 2008. The tuna challenge 2008: Overview and evaluation results. In *Proceedings of the 5th International Conference on Natural Language Generation (INLG'08)*.

K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu. 2001. BLEU: A method for automatic evaluation of machine translation. IBM research report, IBM Research Division.

S. Papineni, T. Roukos, W. Ward, and W. Zhu. 2002. Bleu: a. method for automatic evaluation of machine translation. In *Proc. 40th Annual Meeting of the Association for Computational Linguistics (ACL-02)*, pages 311–318.

R. Ratliff. 1993. Methods for dealing with reaction time outliers. *Psychological Bulletin*, 114(3):510–532.