Introducing Shared Task Evaluation to NLG
The TUNA Shared Task Evaluation Challenges

Albert Gatt[1,2] and Anja Belz[3]

[1] Institute of Linguistics, Centre for Communication Technology, University of Malta
albert.gatt@um.edu.mt
[2] Communication and Cognition, Faculty of Arts, Tilburg University
[3] NLTG, School of Computing and Mathematical Sciences, University of Brighton
asb@bton.ac.uk

Abstract. Shared Task Evaluation Challenges (STECs) have only recently begun in the field of NLG. The TUNA STECs, which focused on Referring Expression Generation (REG), have been part of this development since its inception. This chapter looks back on the experience of organising the three TUNA Challenges, which came to an end in 2009. While we discuss the role of the STECs in yielding a substantial body of research on the REG problem, which has opened new avenues for future research, our main focus is on the role of different evaluation methods in assessing the output quality of REG algorithms, and on the relationship between such methods.

1 Introduction

Evaluation has long been an important topic for Natural Language Processing. Among the developments that have helped to bring it to the forefront of the research agenda of many areas of NLP, two are particularly important. Starting with the work of Spärck Jones and Galliers [1], there has been an attempt to identify the main methodological issues involved, coupled with a more recent drive towards the development of evaluation methods, especially metrics that can be automatically computed [2–4]. A second development has been the organisation of a number of Shared Task Evaluation Challenges (STECs) in areas such as Summarisation[4], Information Retrieval[5] and Machine Translation,[6] which bring together a number of researchers with solutions to a problem based on shared datasets. In addition, several contributions exploring the significance and validity of different evaluation methods, as well as the relationship between them, have appeared in recent years [5–7]. Some of these have discussed the relationship between *intrinsic* evaluation, which evaluates the output of an algorithm in its own right, either against a reference corpus or by eliciting human judgements of quality, and *extrinsic* evaluation, where the system's output is assessed in the

[4] See, for example, http://duc.nist.gov/pubs.html.
[5] See http://trec.nist.gov/.
[6] See http://www.itl.nist.gov/iad/mig/tests/mt.

context of an externally defined task [8, 6, 9]. However, the relationship between results from these two kinds of evaluation remains an open question.

Within NLG, evaluation has typically relied on methods which involve 'humans in the loop'. For example, many intrinsic evaluations have taken the form of elicitation of ratings or responses to questionnaires about various aspects of generated text [10–12]. To take a specific example, the STORYBOOK system [11] was evaluated using questionnaires designed to elicit judgements of quality of different versions of an automatically generated story, obtained by suppressing or including the contribution of different modules to the generation process. Similarly, the SUMTIME weather forecasting system [13] was evaluated by asking participants various questions to assess the extent to which they liked a forecast. The evaluations discussed in the present chapter (see especially Section 4) and elsewhere in this volume [14] also make use of judgements of this kind, among other methods.

Starting with the work of Langkilde [15], automatic intrinsic methods exploiting corpora have mostly been used in evaluations of realisers [16–19], though some corpus-based experiments involving Referring Expressions Generation (REG) algorithms have been reported recently [20–23], including in the TUNA evaluations discussed in the present chapter, and in the GREC shared task competitions [14].

Another evaluation methodology in NLG involves the use of extrinsic, task-based methods to assess the fitness for purpose of an end-to-end generation system within its application domain. Examples include the STOP system [24], which generated smoking cessation letters; BT-45 [25], which produced textual summaries of neonatal intensive care data, and MPIRO [26] and PEACH [27], both of which involved the generation of descriptions of museum artifacts. While such studies tend to be more expensive and labour-intensive, it is often taken for granted within the NLG community that extrinsic methods give a more reliable assessment of a system's utility in doing what it was designed to do.

As the above examples suggest, there is a fairly strong evaluation tradition in NLG. However, *comparative* evaluation is a relatively recent development. This is in part due to the fact that in many cases, researchers develop systems and methods which are not directly comparable, answering to different needs depending on the domain of application. Thus, it is not surprising that comparative studies have either focused on specific subtasks of NLG on which there is significant consensus regarding problem definitions and input/output pairings, or on end-to-end generation tasks. To date, the only example of an evaluation experiment involving directly comparable, independently developed end-to-end generation systems comes from weather forecast generation [28, 29]. As for evaluation of subtasks, the best example is perhaps realisation, for which there is a readily available dataset in the form of the Penn Treebank. Indeed, several groups of researchers have reported results for regenerating the Wall Street Journal corpus [15, 19, 18], though the systems are not directly comparable as they use inputs of different levels of specificity.

There are a number of arguments in favour of comparative evaluation which are worth rehearsing briefly here (they are discussed in greater detail by Belz and Kilgarriff [30]). Perhaps the most obvious immediate benefit is the creation of shared, publicly available data on the basis of which systems can be directly compared. A second argument is that a focus on a common problem speeds up technological progress while conferring additional benefits on the community, such as a growth in size through increased participation. Third, a body of research with results on a single dataset provides baselines against which to compare novel approaches. Finally, given several sets of comparable results from such an evaluation exercise, there is the possibility of follow-up studies, whether these are concerned with identifying properties of systems or algorithms that enhance performance [31], or with methodological issues such as the significance and validity of different evaluation methods [6]. It is with the latter question that the second part of this chapter will be primarily concerned.

The growing interest in comparative evaluation in NLG gave rise to some expressions of interest in the organisation of Shared Tasks, which began to be sounded during discussions at the UCNLG and ENLG workshops in 2005, and continued during a special session on the topic at INLG'06, culminating in an NSF-funded workshop on Shared Tasks and Comparative Evaluation in NLG in April 2007 [32]. At that meeting, various work groups were formed to discuss different aspects of the topic, including the identification of tasks that could potentially form the basis of a STEC. Two of the task types proposed at that meeting – Referring Expression Generation (REG) and Giving Instructions in Virtual Environments (GIVE) – have since been used in a STEC. Two series of REG Shared Tasks have been organised since 2007, one focusing on attribute selection and realisation of references to objects in visual domains (the TUNA STECs, which are the focus of this chapter); the other on generating referring expressions to discourse entities in the context of running text (the Generation of Referring Expressions in Context, or GREC, Tasks; see the chapter by Belz *et al.* [14]). As for instruction giving, the first GIVE Challenge was presented in 2009 (see the chapter by Koller *et al.* [33]). Since 2008, all of these STECs have been presented as part of Generation Challenges, an umbrella event which is now expanding further, with a new Shared Task on Question Generation planned for 2010.[7]

This chapter discusses the first of these series of STECs, those using the TUNA Corpus, which were organised over three years. Our aim is twofold. First, we give an overview of the structure and scope of the Shared Tasks, describing the evaluation procedures as well as the degree of participation and the nature of the contributions (Section 3). Second, we discuss the variety of evaluation methods used and the relationship between them (Section 4). Our findings in this regard have consistently pointed to a problematic relationship between intrinsic and extrinsic methods, echoing findings in other areas of NLP such as Summarisation

[7] See `http://www.nltg.brighton.ac.uk/research/genchal10/` for details of the tasks forming part of Generation Challenges 2010.

[8]. Before turning to these topics, we first give some background on the REG task and the TUNA corpus (Section 2).

2 Background to the TUNA Tasks

Referring Expression Generation (REG) is one of the most intensively studied subtasks of NLG. Like realisation, it has tended to be studied independently of its role within larger NLG systems. Since the seminal work by Appelt, Kronfeld, Dale and Reiter starting in the late 1980s [34–38] widespread consensus has developed regarding the problem that REG algorithms seek to solve, as well as the nature of their inputs and outputs. The problem definition, subscribed to by most traditional approaches to REG (see [39] for a related formulation), is as follows:

Definition 1. *Given a domain of discourse U, consisting of entities and their attributes, and a target referent $r \in U$, find a set of attributes D (the description), such that $[\![D]\!] = \{r\}$.*

This is not to imply that unique identification is *all* that there is to REG; indeed, the foundations laid in the work cited above have been extended in several directions, for example to deal with discourse-anaphoric NPs, issues related to expressiveness and logical completeness (such as dealing with n-place relations, negation and plurality), as well as vagueness. Nevertheless, attribute selection and identification were two central issues in REG and this consensus made it a good candidate for the organisation of a STEC. This was compounded by a growing interest in empirically evaluating REG algorithms, which had hitherto often been justified on the basis of theoretical and psycholinguistic principles, but lacked a sound empirical grounding. Initial forays into empirical evaluation of REG algorithms either used existing corpora such as CocoNut [21, 40], or constructed new datasets based on human experiments [22, 23, 41]. These evaluation studies focused on classic approaches to the REG problem, such as Dale's Full Brevity and Greedy algorithms [37] and Dale and Reiter's Incremental Algorithm [38]. In the studies by Gupta and Stent [21] and by Jordan and Walker [40], these algorithms were evaluated in a dialogue context and were extended with novel features to handle dialogue and/or compared to new frameworks such as Jordan's *Intentional Influences* model [42].

2.1 The TUNA Corpus

One of the datasets constructed as part of the REG evaluations mentioned in the previous section was the TUNA Corpus,[8] a collection of descriptions of objects elicited in an experiment involving human participants. The corpus is made up of several files, each of which pairs a domain (a representation of entities and their attributes) with a human-authored description for one or two entities (the

[8] `http://www.csd.abdn.ac.uk/research/tuna/`

target referent(s)). In each case, there are six further entities in addition to the target(s). In line with the familiar REG terminology, these are referred to as *distractors*.

The descriptions were collected in an online elicitation experiment which was advertised mainly on the University of Zurich Web Experimentation List[9] [43] and in which participation was unrestricted, though data from persons who did not complete the experiment, or who reported a low level of fluency in English, was later removed.

During the experiment, participants were shown visual domains of objects of two types: furniture or people. Domains of the former type consisted of pictures obtained from the Object Databank[10], a set of realistic, digitally created images of familiar objects. These were digitally manipulated to create versions of the same object in different colours and orientations. In the people domain type, the pictures consisted of real black and white photographs of males, which had been used in a previous study by van der Sluis and Krahmer [44]. The attributes of the objects in each domain type are summarised in Table 1. As the table indicates, the two types of images differ in their complexity, in that the photographs of people have a broader range of attributes from which to select. This difference was one of the motivations for eliciting descriptions of two different classes of objects. Apart from attributes such as COLOUR or TYPE, objects were also defined by two numeric attributes (X-DIMENSION and Y-DIMENSION), corresponding to their coordinates in a sparse 3 (row) $\times$ 5 (column) matrix in which the objects were visually presented during the experiment.

Table 1. Attributes and values in the two TUNA domain types.

Attribute	**Possible values**	
	Furniture	*People*
TYPE	*chair, sofa, desk, fan*	*person*
ORIENTATION	*front, back, left, right*	*front, left, right*
X-DIMENSION (column number)	$1, 2, 3, 4, 5$	$1, 2, 3, 4, 5$
Y-DIMENSION (row number)	$1, 2, 3$	$1, 2, 3$
SIZE	*large, small*	-
COLOUR	*blue, red, green, grey*	-
AGE	-	*young, old*
BEARD	-	0 (`false`), 1 (`true`)
HAIR-COLOUR	-	*dark, light, other*
HASHAIR	-	0 (`false`), 1 (`true`)
HASGLASSES	-	0 (`false`), 1 (`true`)
HASSHIRT	-	0 (`false`), 1 (`true`)
HASTIE	-	0 (`false`), 1 (`true`)
HASSUIT	-	0 (`false`), 1 (`true`)

[9] `http://genpsylab-wexlist.unizh.ch`

[10] The Object Databank was constructed by Michael Tarr and colleagues and is available on `http://alpha.cog.brown.edu:8200/stimuli/objects/objectdatabank.zip/view`.

Design A full description of the design of the experiment can be found in Gatt *et al.* [23]. Here, we focus primarily on those aspects of the design which are relevant to the STECs. The main within-participants condition 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 a small number of participants nevertheless included this information. The corpus annotation indicates which of the two conditions a description was elicited in (see 2 and the description of the annotation below). The other within-participants variable manipulated in the experimental design concerned the number of entities referred to. One third of the trials were singular, that is, there was one target referent and six distractor objects. The rest were plural, with two target referents, which were SIMILAR one-third of the time (that is, had identical values on their attributes[11]), and DISSIMILAR in the rest of the trials. Since plurals did not feature in any of the TUNA STECs, we shall not discuss the distinction any further here.

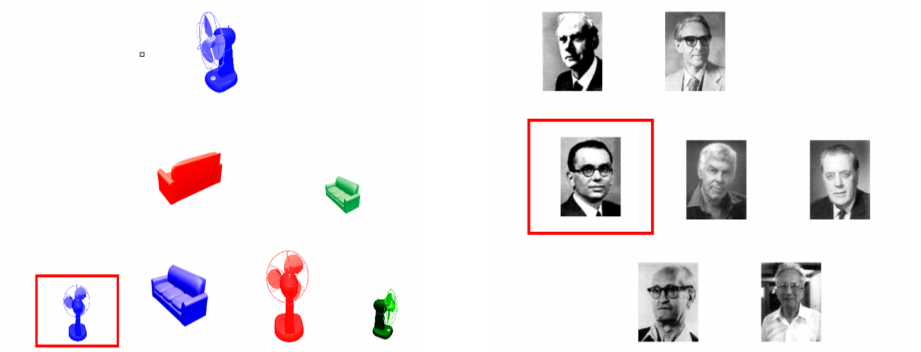
The experimental design also contained a further between-groups condition, which manipulated whether the communicative situation in which descriptions were produced was perceived as 'fault-critical' or not (±FC). The way this was done is explained below. However, for the TUNA STECs, data from the ±FC conditions was merged. There were two reasons for this. First, distinguishing (±FC) descriptions would have complicated the design of the shared tasks, taking them beyond the conventional task definition given above. Second, preliminary analysis of the corpus suggested that this variable did not significantly affect the content of descriptions. However, it is worth noting that other researchers have shown an impact of communicative intent and 'importance' on descriptions produced by humans [45, 46].

The experiment consisted of a total of 38 trials, each consisting of a visual domain for which a participant typed a description for the target referent. The trials were balanced in the following sense. For each possible subset of the attributes defined for a given domain type (shown in Table 1), there was an equal number of trials in which an identifying description of the target(s) required the use of all those attributes (i.e. leaving one of the attributes out would have resulted in an unsuccessful identifying description) *unless* it included locative information. For example, there was a furniture trial in which a target could be distinguished by using COLOUR and SIZE. TYPE was never included in this calculation because previous research has suggested that this attribute is included by human speakers irrespective of whether it has any discriminatory value in an identification task [47]. The locative attributes (X-DIMENSION and Y-DIMENSION) were not balanced in the same manner, because the position of objects in the

[11] The exception was TYPE in the furniture domain: SIMILAR plurals in this domain were identical on all their visual attributes, including COLOUR and ORIENTATION, but had different TYPE values. For example, one object could be a chair and the other a sofa. This does not apply in the people domain, since all entities were of the same type.

grid in which they were presented was determined randomly at runtime for each trial and each participant.

Procedure All participants were exposed to all 38 trials in randomised order. Examples of trials with both people and furniture are shown in Figure 1. A trial consisted of a visual domain consisting of one or two target referents and six distractor entities in a sparse 3 × 5 grid. The target(s) were indicated by a red border. Participants typed a description of the referent in a text box in answer to the question: *Which object is surrounded by a red border?*



(a) Furniture trial (images by Michael Tarr and colleagues, distributed with the Object Databank).

(b) People trial (images from a previous study by van der Sluis and Krahmer [44]).

Fig. 1. Trials in the TUNA elicitation experiment.

Participants were told that their descriptions were being used to test a language understanding system in real time. Following each trial, some objects were removed from the screen and participants were given feedback as to whether the 'system' had managed to interpret their description correctly. In actual fact, the objects to be removed were determined in advance, so that the 'system' was successful 75% of the time. As noted above, one of the conditions manipulated in the experiment was ±FC: in the −FC condition, participants were told that, should the system make an error, they could correct it (by clicking on the right target objects). In the −FC condition, this was not possible.

Participation and corpus size Sixty participants completed the experiment on a voluntary basis. All were self-reported native or fluent speakers of English. This gave rise to (60 × 38 =) 2280 descriptions paired with their domain representation, of which 780 were singular (420 furniture and 360 people descriptions), and 1500 were plural (780 furniture descriptions and 720 people descriptions). Only the singular sub-corpus was used for the TUNA STECs.

```
<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>
```

Fig. 2. XML format of corpus items.

Annotation The XML format used in the TUNA STECs, shown in Figure 2, is a variant of the original format of the TUNA corpus.[12] The root `TRIAL` node has a unique ID and an indication of the ±LOC experimental condition. The `DOMAIN` node contains 7 `ENTITY` nodes (1 target and 6 distractors), which themselves contain a number of `ATTRIBUTE` nodes defining the possible properties of an entity in attribute-value notation. The attributes cover all properties defined in Table 1, 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 corresponding image that was used in the experiment. 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 this set of attributes only.

[12] The annotation manual for the original TUNA Corpus is available from the corpus website.

3 The TUNA Challenges 2007-2009

Starting in 2007, three Shared Task Evaluation Challenges (STECs) were organised using the TUNA corpus.[13] In what follows, we refer to these as TUNA'07, TUNA'08, and TUNA'09.[14] TUNA'07 [48] was presented five months after the NSF workshop, in September 2007 at the UCNLG+MT workshop at MT Summit XI in Copenhagen. It was organised in the spirit of a pilot event to gauge community interest in STECs, with encouraging results. This was followed by TUNA'08 [49], held as part of Generation Challenges 2008, which also included the first GREC task. The results were presented during a special session at INLG'08 in Salt Fork, Ohio. The third and final edition, TUNA'09 [50], was presented at ENLG'09 in Athens as part of Generation Challenges 2009, which also included two GREC tasks and the first GIVE challenge.

In addition to the shared tasks proper, all of these events included two special tracks: (i) an open submission track in which participants could submit any work involving the data from any of the shared tasks, while opting out of the competitive element, and (ii) an evaluation track, in which proposals for new evaluation methods for the shared task could be submitted. Generation Challenges 2009 additionally offered a task proposal track inviting proposals for new shared tasks.

The idea behind the Open Track was partly motivated by discussions at the INLG'06 Special Session and the NSF Workshop, in which one of the reservations expressed in relation to STECs was that they might result in a narrowing of the scope of research on a topic by emphasising a small set of tasks at the risk of stifling novel ideas. The Evaluation Methods track was intended to complement the growing interest within the NLP community in evaluation methods mentioned in Section 1, especially in view of the cautionary notes sounded from various quarters concerning the validity of certain metrics and the relationship between them. In the event, there was only one submission to the Evaluation Methods track in ASGRE'07 [51], and no submissions in subsequent editions. Similarly, the Open Track attracted one submission during TUNA-REG'08, proposing a method for exploiting TUNA `ATTRIBUTE-SET`s to generate descriptions in Brazilian Portugese [52].

In the remainder of this section, we give an overview of the three TUNA STECs, focusing on their structure and the evaluation methods used. We also briefly discuss the level of participation in the various sub-tasks of each STEC. An exhaustive description of all the different submissions is beyond the scope of this chapter; the relevant publications by participants can be found in the proceedings of the special sessions of the events at which they were presented.

[13] The Participants' Packs from all three years, including evaluation tools and documentation, are available from http://www.nltg.brighton.ac.uk/home/Anja.Belz

[14] Though these were not the original titles of the STECs, we have adopted this naming convention for ease of reference throughout this chapter.

3.1 Structure and Scope of the TUNA Shared Tasks

The structure of the Shared Tasks is summarised in Table 2. As the Table indicates, over the three editions of the TUNA STECs, there was a gradual broadening of the scope of the REG tasks involved.

Table 2. Summary of tasks in the three editions of the TUNA Challenges.

Edition	Tasks	Attribute Selection	Realisation
TUNA'07	TUNA-AS	y	n
TUNA'08	TUNA-AS	y	n
	TUNA-R	n	y
	TUNA-REG	y	y
TUNA'09	TUNA-REG	y	y

TUNA'07 consisted of a attribute selection single task, TUNA-AS. Although the task definition was close to the one traditionally used in the REG literature and given above in Definition 1, no emphasis was made on unique identification in order to avoid narrowing its scope. (As it turned out, all peer systems submitted to the STEC performed unique identification.) In the context of the TUNA Corpus, the input to peer systems was a `DOMAIN` node, and the output was an `ATTRIBUTE-SET` which represented the selected attributes for the designated target referent in the `DOMAIN`. The scope of the Shared Tasks was extended in TUNA'08, which kept the TUNA-AS task from the 2007 edition, but included two others, TUNA-R and TUNA-REG, both of which included realisation. Thus, the output of systems in the latter two tasks was no longer an `ATTRIBUTE-SET`, but a `WORD-STRING`. In TUNA-R, which focused exclusively on realisation, the task was defined as mapping from an input `ATTRIBUTE-SET` to an actual Natural Language description, while TUNA-REG was defined as an end-to-end referring expression generation task, which went from an input `DOMAIN` to an output `WORD-STRING`, combining content determination (the TUNA-AS Task) and realisation (the TUNA-R Task). However, no constraints were imposed on the way this was to be achieved (for example, content determination and realisation could be implemented as separate modules in a pipeline or interleaved). This broadening of scope resulted in several submissions which presented novel approaches to realisation in addition to content determination. Finally, TUNA'09 was organised as a follow-up to TUNA'08 and consisted of the single end-to-end TUNA-REG task. The aim was to provide participants of TUNA'08 with the possibility of improving on the results obtained in the 2008 edition, while also including new participants.

In all editions, there was a three-stage participation procedure. Potential participants first registered to express their interest in participating. Shortly after, a participants' pack consisting of training and development data (see Section 3.2 below), software to compute some of the relevant evaluation measures (see Section 3.4), and detailed documentation for both data and software, was made available. Participating teams were then required to submit a writeup describ-

ing their system(s), including evaluation scores computed on the development set with evaluation software provided by us. Once this was submitted, they could download the test data inputs and submit their system outputsfor the test data, after which the organisers computed the evaluation scores on the test data outputs.

3.2 Datasets

In each edition of the TUNA tasks, participants were provided with both training and development datasets. Unlike the test data, which was provided to participants in the form of inputs only, both the training and development sets consisted of paired inputs and outputs. Although no restriction was placed on the way training and development data were to be used in the process of designing a system, the purpose of the division was twofold. First, participants could conduct an iterative training and testing process in the course of building their systems, using the development dataset for the latter purpose. Second, the development data was also used by participants to report scores on a subset of the evaluation methods used for a given task. These scores, computed on 'seen' data, were then compared by the organisers to the final evaluation scores obtained on the test data outputs.

For the TUNA'07 edition, test data as well as training and development data came from the original TUNA corpus, which had not been publicly released at that point. The corpus was divided into 60% training, 20% development and 20% test data. For TUNA'08, the original corpus data was divided into 80% training and 20% development data, and two new test sets were constructed by partially replicating the original TUNA experiment described in Section 2.1. One of the new test sets was used for the TUNA-AS and TUNA-REG tasks in 2008, and later re-used in the 2009 replication of TUNA-REG; the other was used exclusively for TUNA-R in 2008.

The experiment to construct the new test sets was designed to ensure that each `DOMAIN` had two reference outputs. Thus, the corpus-based evaluations for the 2008 and 2009 STECs were conducted against multiple descriptions for each input `DOMAIN`. Like the original TUNA experiment, the new one was conducted over the web, and advertised mainly among staff and students of the Universities of Aberdeen and Brighton. The task given to experimental participants was identical to the original, except for the following differences. First, no plurals were included, since the TUNA tasks only focused on singular descriptions. Second, the new experiment did not include the 'interactive' component of the original, whereby participants were told that they were interacting with a language understanding system. This is because we decided not to replicate the ±FC manipulation, given that data from the +FC and −FC conditions had been collapsed for the purposes of the TUNA tasks. The experiment was completed by 218 participants, of whom 148 were native speakers of English. The test sets were constructed by randomly sampling from the data gathered in the experiment from native speakers only: both sets consisted of 112 items, divided equally into

furniture and people descriptions and sampled evenly from both experimental conditions (±LOC).

3.3 Participation

Over the three editions, the various tasks in the TUNA STECs attracted a total of 61 different systems from 10 different teams, based in 10 countries and 5 continents. This turnout suggests that exercises in comparative evaluation are positively viewed on the whole within the NLG community. On the other hand, participation was no doubt enhanced by the fact that the REG task is very familiar to many members of the community (this was the reason why it appeared to be a good candidate for a first try at an NLG STEC).[15] While TUNA'07 had 22 systems from 6 teams, and the three tasks in TUNA'08 had a total of 33 systems from 8 teams, TUNA'09 had 6 systems from 4 teams. These participation levels reflect the different numbers of tasks on offer, but may also be an indication that the TUNA tasks had reached a saturation point after three years.

On the whole, these events yielded a substantial body of new research on the REG problem. Many of the proposed solutions to the attribute selection problem built on classic approaches such as Dale and Reiter's Incremental Algorithm [38] or Dale's Full Brevity Algorithm [37]; or on well-understood formalisms such as graphs [53]. However, there were also a number of new algorithmic approaches, perhaps the most notable of which were those using evolutionary techniques [54, 55].

In general, the various approaches to attribute selection in the TUNA-AS task had a markedly empirical orientation, relying on corpus-based frequencies to drive attribute selection (e.g. [56, 57]). This was the case even when the algorithms were extensions of existing frameworks. For instance, Theune *et al.* [58] used the graph-based approach to REG proposed by Krahmer *et al.* [53], in which a REG domain is represented as a graph whose nodes are entities and whose edges represent properties. Theune *et al.* used cost functions derived from corpus data to weight graph edges, thus making some attributes more likely to be selected than others.

There was also a notable interest in using individual profiles, that is, speaker-based corpus-derived statistics to guide selection. Thus, Bohnet [59, 60] proposed a content determination algorithm that combined well-known heuristics (such as Full Brevity) with individual author identities to prioritise attributes during selection. A similar rationale was followed by di Fabbrizio *et al.* [61, 62], who showed that using frequencies from individual speakers to guide attribute selection results in improvements on intrinsic evaluation scores. These approaches exploited the fact that the corpus contains several descriptions collected from different authors in different experimental trials.

TUNA'08 was also the point where realisation was introduced, and this resulted in an increased focus on realisation methods, which are often ignored in

[15] Tasks which began to be organised more recently, namely GREC and GIVE, have so far attracted fewer submissions, and this may be due to their relative novelty.

the REG literature. This too yielded a broad spectrum of proposals, again with a decidedly empirical flavour, for example relying on corpus-derived templates or language models [61], as well as case-based reasoning [63, 55].

The breadth of approaches to both attribute selection and realisation may serve to open up new avenues of research in future work on REG. Moreover, the various TUNA tasks have now yielded a sizeable body of evaluation data which can act as a baseline against which to compare new systems, but also as the underlying dataset for 'meta-analyses' of various kinds, for example comparing different algorithmic approaches (see [31] for an example based on the TUNA'08 data) or different evaluation methods (see Section 4).

3.4 Evaluation Methods

Perhaps the most important aspect of the TUNA STECs was their use of a broad variety of evaluation methods. These included both intrinsic methods, which ranged from automatically computed metrics to methods involving human participants (Table 3), and extrinsic methods involving humans (Table 4).

Table 3. Intrinsic evaluation methods used in the three editions of the TUNA tasks.

Criterion	Method	Type	Tasks
Humanlikeness	DICE	automatic	TUNA-AS'07-8
	MASI	automatic	TUNA-AS'08
	ACCURACY	automatic	all tasks
	BLEU	automatic	TUNA-R, TUNA-REG
	NIST	automatic	TUNA-R, TUNA-REG
	EDIT	automatic	TUNA-R, TUNA-REG
Theoretical	MINIMALITY	automatic	TUNA-AS
	UNIQUENESS	automatic	TUNA-AS
Judged appropriateness	ADEQUACY	human	TUNA-REG'09
	FLUENCY	human	TUNA-REG'09

Table 4. Extrinsic evaluation methods used in the three editions of the TUNA tasks.

Criterion	Method	Type	Tasks
Task effectiveness	Identification Time	human	TUNA-AS'07, TUNA-REG
	Identification Accuracy	human	TUNA-AS'07, TUNA-REG
Ease of comprehension	Reading Time	human	TUNA-AS'07, TUNA-REG'08

The decision to use a variety of methods was motivated by the experience of other shared task evaluation challenges, which has shown that the use of a single method of evaluating participating systems can cause loss of trust and/or loss of interest, in particular if the method is seen as biased in favour of a particular

type of system (a criticism that has been levelled at BLEU in Machine Translation research [5]) or if it severely restricts the definition of quality (as in the case, for example, of the PARSEVAL metric in syntactic parsing). An additional motivation, already hinted at in Section 1, came from recent work suggesting that different evaluation methods can yield very different results and that the relationship between them is often not straightforward, making comparisons based on multiple methods even more desirable. The particular choice of methods was informed by the following criteria.

Humanlikeness: It is more or less taken as read in many areas of NLP evaluation that the greater the similarity between a system's output and a particular corpus of human-produced reference outputs, the better the system. This is the rationale behind metrics such as BLEU, ROUGE and NIST. The TUNA Challenges used a range of methods that assessed humanlikeness. For the TUNA-AS task, this was defined in terms of set-theoretic metrics comparing attribute sets produced by systems to those produced by human authors on the same `DOMAIN`. We used the Dice coefficient in the 2007 version of the task, adding the MASI score [64] in 2008. Both metrics measure the degree of overlap between sets. Let A and B be two sets of attributes, corresponding to two descriptions. Dice, which is computed as shown in (1), ranges between 0 and 1, where 1 indicates identity.

$$Dice(A, B) = \frac{2 \times |A \cap B|}{|A| + |B|} \tag{1}$$

The MASI score, which is based on the Jacard coefficient, is also a set comparison metric that ranges between 0 and 1:

$$MASI(A, B) = \delta \times \frac{|A \cap B|}{|A \cup B|} \tag{2}$$

where δ is a *monotonicity coefficient* defined as follows:

$$\delta = \begin{cases} 0 \text{ if } A \cap B = \emptyset \\ 1 \text{ if } A = B \\ \frac{2}{3} \text{ if } A \subset B \text{ or } B \subset A \\ \frac{1}{3} \text{ otherwise} \end{cases} \tag{3}$$

Dice and MASI differ in two respects. First, the Jacard coefficient in MASI tends to penalise a low degree of overlap between sets more than Dice. Second, the monotonicity coefficient (3) biases the score in favour of those cases where one set is a subset of the other.

In the TUNA-R and TUNA-REG tasks, where system outputs were `WORD-STRING`s, we used Levenshtein (String Edit) Distance, as well as BLEU and NIST. Edit Distance computes the cost of transforming one string into another based on the number of insertions, deletions and substitutions. In the TUNA STECs, the cost of insertions and deletions was set to 1, while substitutions had a cost of 2. Levenshtein distance has a value of 0 when the strings are identical, and its upper bound depends on the length of the longest string in the comparison pair. BLEU

and NIST are n-gram based aggregate scores. For BLEU, we set the parameter that controls the maximum length of n-grams taken into account to the standard value of 4 in TUNA'08. However, this gave odd results due to the brevity of the strings being compared. Hence, we used a maximum n-gram length of 3 in TUNA'09. The primary difference between BLEU and NIST is that the latter gives more importance to less frequent n-grams.

Finally, all tasks used a measure of ACCURACY, which computed the proportion of times a system output was identical to a corpus instance. The precise definition of this depends on the task: in the case of TUNA-AS, the item of comparison was the `ATTRIBUTE-SET`, while for the TUNA-R and TUNA-REG tasks, it was the `WORD-STRING`.

Theoretically motivated criteria: In the case of REG, the Shared Tasks also provided an opportunity to evaluate systems in terms of some assumptions that have dominated the field since the work of Appelt [34] and Dale and Reiter [38]. The first measure, UNIQUENESS, was motivated by the emphasis in the REG literature on *identification* (see Definition 1). It was defined as the proportion of `ATTRIBUTE-SET`s produced by a system which uniquely identified the intended referent. All systems scored 100% on UNIQUENESS; therefore, we shall not discuss it further. In addition to identification, a lot of research on REG has also revolved around a (theoretical) interpretation of the Gricean maxims [65], particularly the Maxim of Quantity, which was interpreted as a constraint on REG algorithms to include no more information in a description than absolutely required for identification (see Section 4.1 for further discussion). This criterion was straightforwardly applicable in the TUNA-AS: the MINIMALITY measure computed the proportion of attribute sets produced by a system that (a) uniquely identified their target referent and (b) were minimal, in the sense that any alternative attribute set that distinguished the target referent within the domain would be at least as long as the one returned by the system.

Judged appropriateness: While automatically computed metrics compare system outputs against a gold standard corpus, an alternative way to assess intrinsic quality is to obtain human judgements, a method which is frequently employed in NLG. Human judgements were included in the final edition of the TUNA STECs, in an experiment that included both system outputs and the human-authored reference descriptions from the corpus. The experiment involved 8 native speakers of English who were doing a Masters degree in a linguistics-related subject, recruited from among post-graduate students at UCL and the Universities of Sussex and Brighton. Participants were presented with a description of a target referent, together with a visual representation of the input `DOMAIN` (see Figure 1). They were asked to rate the description based on two criteria, ADEQUACY and FLUENCY. The following are the relevant excerpts from the instructions given to them:

1. (ADEQUACY) 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. (FLUENCY) How fluent is this description? Here your task is to judge how well the phrase reads. Is it good, clear English?

In addition, the instructions to participants emphasised that the two measures are distinct, that is, it is perfectly possible for a description to be highly adequate but non-fluent. It appears that participants were able to assess the two criteria independently to some extent: subsequent analysis revealed a correlation of just 0.6 between the two measures which was not statistically significant [50].

For both kinds of judgements, ratings were obtained using a slider, whose position returned an integer between 1 (worst) and 100 (best). Participants were not shown the actual integer values, but only manipulated the slider's position. For each trial, the slider was initially placed in the middle. The motivation for using a slider rather than the more standard rating scale (e.g. a scale from 1 to 5) was that means obtained from such ordinal scales tend to be hard to interpret. This is partly because participants may attach different meanings to the values on the scale and the distances between them (for example, interpreting the distance between 1 and 2 and between 4 and 5 differently); moreover, participants who provide the ratings in the first place have no opportunity to rate an item as falling within an interval. The use of parametric methods to analyse this type of ordinal data is not generally considered appropriate. The method used here, by contrast, relied exclusively on spatial distance in a visual modality: it was the degree of displacement of the slider that reflected participants' judgements since, although the slider position mapped to a numeric value, the value itself was not known to participants. Our assumption was that the physical position of a slider would not be as susceptible to differences in 'meaning' as would a value on an ordinal scale.

Our use of sliders was partially inspired by studies using *Magnitude Estimation* (ME) [66], which requires participants to make judgements by comparing an item to a modulus (i.e. an item that serves as a comparison point throughout the experiment) using a numeric scale of their choice. In one version of this task, participants carry out judgements in a visual modality rather than using numbers, e.g. by indicating the degree of acceptability of an item relative to the modulus using the length of a line. Bard *et al.* [66] have shown that participants are remarkably self-consistent, with high correlations obtained between grammaticality judgements in the numeric and visual modalities. Similarly, Gatt and van Deemter [67] asked participants to rate the plausibility of plural referring expressions by comparing them to a modulus using both numeric scales and sliders. This experiment, which focused on plausibility rather than grammaticality, also found very high degrees of self-consistency in the ratings given by participants.

Task effectiveness and comprehension: All the TUNA STECs involved a task-based (extrinsic) evaluation which was based on the task definition typically associated with REG (Definition 1), that is, the idea that the output of REG algorithms

aims to help a reader or hearer identify the intended referent in the relevant domain of discourse. The ability of algorithms to do this was tested in a series of experiments in which participants were exposed to a description (produced either by a system or by a human author), and were shown the visual domain. The task was to identify the intended referent based on the description, by clicking on the corresponding picture. The relevant dependent variables were the speed with which a referent was identified (identification time), and the proportion of correct identifications per system (identification accuracy). In addition, reading time was a relevant variable in one of the experiments. Over the course of the three editions, we experimented with different methods, which differed primarily in whether or not the presentation of the description was made at the same time as the presentation of the visual domain, and in whether the description was read or heard:

1. In the experiment carried out as part of TUNA'07, a trial consisted of a visual domain and a written description presented on screen in tandem. Thus, in this experiment, reading/comprehension time and identification time could not be distinguished. As a result, a follow-up experiment was performed after the Shared Task [6], which separated reading/comprehension and identification. This method, which was also used for TUNA'08, is described below. Since in this task submissions consisted of `ATTRIBUTE-SET`s, all submitted outputs were realised using a very simple template-based realiser (written by Irene Langkilde-Geary, Brighton University, for this purpose) which always realises each attribute in the same way and in the same position regardless of context, except that it disambiguates negated attributes by always realising them at the end of the description.[16]
2. In TUNA'08, as well as in the follow-up identification experiment for TUNA'07 referred to in (1) above, reading/comprehension and identification were separated, by splitting each trial so that a description was first presented and read by a participant, and the visual domain was subsequently presented once a button was pressed. The first phase of a trial was therefore a self-paced reading task, while the second was the identification task proper. This is why *ease of comprehension*, operationally defined in terms of reading time, emerged as a separate evaluation criterion in this experiment (see Table 4). In this edition, only systems submitted to the end-to-end task (that is, TUNA-REG) were included in the evaluation experiment, obviating the need to use a realiser to render attribute sets as strings.
3. In TUNA'09, reading was eliminated in favour of a paradigm in which a description was heard over a headset *while* a domain was visually presented. This method was adopted because it seemed to better approximate realistic comprehension settings; there are psycholinguistic findings to the effect that comprehension of spoken language descriptions and search/identification are tightly coupled, with identification occurring incrementally as a description

[16] For example, the realiser generates *the person with glasses and no beard* rather than *the person with no beard and glasses*, which is potentially ambiguous.

is understood [68]. In the present case, we used a speech synthesiser to read system-produced and human-authored descriptions.[17]

4 Comparing Different Evaluation Methods

Having described the evaluation methods used and the criteria motivating them, we now turn to a comparison of the different classes of methods. In what follows, we identify points of correspondence and divergence between the various methods using Pearson's r product-moment correlations. All analyses reported below were carried out by comparing means of different systems on the relevant measures. All significance tests are two-tailed. In the following subsections, we first report the results and then discuss some of their implications in Section 4.4.

4.1 Minimality Versus Humanlikeness

In the REG literature, the role of minimality (or 'brevity'; [37]) received prominence following the work of Dale and Reiter [38]. In that paper, various computational interpretations of the Gricean Maxim of Quantity were discussed. Echoing an earlier observation by Appelt [34], the authors showed that a literal interpretation of this maxim (along the lines of the MINIMALITY criterion used in TUNA-AS) led to an exponential running time for REG algorithms, since exhaustive search needs to be performed in order to identify the smallest possible set of attributes that identifies an entity. The authors proposed the Incremental Algorithm as a compromise, since it does not perform backtracking, performing a hillclimbing search along a predefined order of attributes, and halting once a referent has been distinguished. This algorithm does not guarantee that the solution returned is the briefest available, but the authors argued that this is in fact a psycholinguistically plausible outcome, since speakers appear to produce overspecified descriptions in referential tasks. In fact, psycholinguistic results to this effect have been consistently reported for some time [47, 70, 71]. Nevertheless, brevity has remained a central concern of many recent approaches to REG. For example, Gardent [72] has proposed a constraint-based approach to generating minimal plural descriptions, in order to avoid the excessively complex outputs that are produced by a generalisation of the Incremental Algorithm proposed by van Deemter to handle negation and disjunction [73].

Using the results from the 2007 and 2008 editions of the TUNA-AS task, we computed the correlation between the MINIMALITY scores (the proportion of minimal descriptions produced by a system) and the humanlikeness of the `ATTRIBUTE-SET`s, as reflected by the DICE and MASI coefficients.[18] For both tasks,

[17] We used the University of Edinburgh's Festival speech generation system [69] in combination with the nitech_us_slt_arctic_hts voice, a high-quality female American voice.

[18] For the TUNA-AS'07 task, MASI was calculated after the Shared Task, as it was not originally included as an evaluation measure; see Table 3.

Table 5. Pearson's r correlations between MINIMALITY and humanlikeness as measured by the Dice and MASI coefficients. **: significant at $p \leq .001$; * $p \leq .05$.

Task	DICE	MASI
TUNA-AS 2007	-.90**	-.790**
TUNA-AS 2008	-.96*	-.90*

there was a strong negative correlation between MINIMALITY and both humanlikeness scores, as shown in Table 5, indicating that systems which produced more minimal outputs were overall less likely to match corpus descriptions. Given the robust psycholinguistic findings showing that people overspecify descriptions, these results are relatively unsurprising.

We also looked at the correlation between minimality and identification time in the TUNA-AS'07 data, using the results from the follow-up identification experiment reported in Gatt and Belz [31] and described in the previous section (see also Section 4.3 below). Minimality did not correlate significantly with Reading Time ($r = .18; p > .05$) or with Identification Accuracy, the proportion of identification errors made by experimental participants ($r = .51; p > .05$). However, we found a significant *positive* correlation between Identification Time and Minimality ($r = .56; p < .05$), suggesting that minimal descriptions slowed down identification. However, a caveat needs to be raised. Minimal descriptions in the present study typically consisted of a set of attributes such as COLOUR and SIZE, with no TYPE information, because TYPE was never required to distinguish an object in the TUNA domains. For the purposes of the experiment, such descriptions were rendered by the realiser using a default head noun to realise the TYPE attribute. Furthermore some attributes were realised using several words, others as single words. As a result, minimal attribute sets could conceivably have been realised as longer word strings than non-minimal ones.

4.2 Measures of Humanlikeness Versus Task Effectiveness

The comparison of humanlikeness and task effectiveness in this section is based on results from all three editions of the TUNA STECs. We make the following comparisons:

1. For the TUNA-AS'07 task, we use the results of the *follow-up* identification experiment. Unlike the original experiment reported for that STEC, the follow-up utilised the same experimental paradigm as for TUNA'08, separating reading time and identification time, as well as computing identification accuracy (see Section 3). We report correlations for these measures with DICE and MASI scores on attribute sets. In addition, we also report correlations with the *string* metrics used in subsequent STECs, namely EDIT, BLEU and NIST. These were computed as part of our follow-up evaluation, using the strings generated by the template-based realiser described in Section 3.4.
2. For the 2008 and 2009 tasks, we report correlations between the extrinsic measures and string-based metrics only (i.e. not DICE or MASI). This is

because only the TUNA-REG submissions were included in the extrinsic evaluation for the 2008 edition (to the exclusion of the TUNA-AS and TUNA-R systems), while the 2009 edition consisted only of the TUNA-REG task. In both cases, participants were not required to submit `ATTRIBUTE-SET`s (only `WORD-STRING`s); hence, computation of DICE and MASI is not possible. Note that reading time is not available for TUNA'09, since descriptions were presented as auditory stimuli.

All correlations are displayed in Table 6.

Table 6. Pearson's r correlations between intrinsic and extrinsic measures. RT: Reading Time; IT: Identification Time; IA: Identification Accuracy; *: significant at $p \leq 0.05$.

Measure	**Task**	RT	IT	IA
DICE	TUNA-AS (2007)	0.12	-0.28	-0.39
MASI	TUNA-AS (2007)	0.23	-0.17	-0.29
EDIT	TUNA-AS (2007)	-0.30	0.09	0.22
	TUNA-REG (2008)	0.1	0.09	0.22
	TUNA-REG (2009)	-	0.68	-0.01
BLEU	TUNA-AS (2007)	0.39	0.04	-0.08
	TUNA-REG (2008)	0.22	0.04	-0.37
	TUNA-REG (2009)	-	-0.51	0.49
NIST	TUNA-AS (2007)	0.54^*	0.22	0.03
	TUNA-REG (2008)	0.54	0.27	-0.45
	TUNA-REG (2009)	-	0.06	0.60

The most striking feature of the results in Table 6 is that none of the automatically computed metrics that assess humanlikeness have a significant correlation with any of the task-performance measures. The one exception to this trend is the significant (though not very strong) positive correlation between NIST and reading time in the TUNA-AS'07 results. In general, there is no evidence of a relationship between humanlikeness as measured by the metrics used here, and task performance defined in terms of reading speed, identification speed and identification accuracy. On the other hand, separate analyses showed strong correlations *within* the two classes of measures, that is, systematic covariation is observed between different extrinsic measures on the one hand, and intrinsic ones on the other (for details, see the TUNA Shared Task reports [48–50]).

There have been some other publications reporting a similar lack of correlation between intrinsic and extrinsic evaluation results [8, 74], as well as experimental studies showing divergent trends between human judgements and/or preferences and actual task performance [71, 75]. However, our automatic intrinsic scores also do not correlate at all with the *human* intrinsic scores, and this finding is not supported by a set of similar results in the literature (quite the contrary; see next section). At the same time, the human intrinsic scores do correlate with extrinsic results very straightforwardly.

So we have a situation where (surprisingly) human and automatic intrinsic results do not correlate, and automatic intrinsic and extrinsic results do not correlate either, whereas human intrinsic and extrinsic results do correlate. If human and automatic intrinsic results did correlate (as they should), then (because human intrinsic and extrinsic correlate) one might also find a correlation between automatic intrinsic and extrinsic results.

We return to this set of interrelated issues in the next section, where we discuss a possible explanation.

4.3 Human Intrinsic Measures Versus Automatic Intrinsic and Human Extrinsic Measures

Finally, we turn to the relationship between judged appropriateness, on the one hand, and extrinsic and automatically computed intrinsic measures, on the other. In this case, we focus exclusively on the results from TUNA-REG'09, since this is the only task for which judged appropriateness ratings (FLUENCY and ADEQUACY) were collected. The correlations between the judgements and the extrinsic and automatic intrinsic measures are displayed in Table 7.

Table 7. Correlations between judged appropriateness and extrinsic measures. IT: Identification Time; IA: Identification Accuracy; $^{**}: p \leq .001$; $^{*}: p \leq .05$.

Measure	IT	IA	EDIT	BLEU	NIST
FLUENCY	-0.89*	0.50	-0.57	0.66	0.30
ADEQUACY	-0.65	0.95**	-0.29	0.60	0.48

The table shows no significant correlation between any of the automatically computed intrinsic measures and the two kinds of ratings obtained from participants However, some of the correlations between the ratings and the extrinsic measures are significant. In particular, there is a strong negative correlation between FLUENCY and identification time, suggesting that descriptions judged as highly fluent led to faster identification. On the other hand, there was no significant correlation between FLUENCY and identification accuracy. The relationship between ADEQUACY and identification time goes in the same direction as that between FLUENCY and identification time, but fails to reach significance. In contrast, the correlation of ADEQUACY with identification accuracy is very strong and highly significant. These correlations are exactly as one would expect.

In other NLP subfields – MT and summarisation in particular – judged appropriateness ratings of fluency and adequacy are used as a kind of gold standard against which to measure automatic metrics. In other words, the higher its correlation with the human ratings, the better an automatic metric is deemed to be. In those fields, very strong correlations with human judged appropriateness ratings are typically reported for BLEU (and its NIST variant). In NLG too, correlations above .8 have been reported for weather forecast texts [28]. Closest to the present task, the GREC evaluations have also revealed strong and highly

significant correlations between automatic intrinsic metrics and fluency ratings in particular, for both tasks involved [14].

So how are we to interpret the complete lack of significant correlations between our intrinsic automatic and intrinsic human-assessed scores? Is the TUNA Task so different not just from MT, SUMMARISATION and data-to-text generation, but even from the GREC Tasks (which are also REG tasks), that the same metrics that usually achieve high correlations with human judgements for these tasks simply do not work for referring expression generation from attribute-based representations of sets of entities?

As this seems an unconvincing proposition, we need to look for further explanation. A methodological issue that might have affected the results reported in this section concerns the nature of the reference data against which automatic scores are computed. The TUNA Corpus was constructed through an experiment which, being web-based, was relatively unrestricted. The corpus contains a high degree of variation among individual authors, with some adopting a highly telegraphic style, while others produced full NPs of the sort that are typically exemplified in the REG literature.[19] Some variation may also be due to the fact that some of the people who wrote the REs in the corpus were not native speakers.

We have some insight into the perceived quality of these descriptions, as some of them were included in our TUNA'09 judged appropriateness experiments. In these experiments we asked participants to evaluate the competing systems' outputs for the test data inputs, so the human-produced descriptions we included in the experiments were the corresponding descriptions in the corpus (the test data reference 'outputs'). Because we had two descriptions for each input in the corpus, we randomly split them into two groups of single outputs, HUMAN-1 and HUMAN-2. Despite the fact that the TUNA'08/09 test set was obtained in a more controlled setting than the original TUNA data (see Section 3.2 above), HUMAN-1 and HUMAN-2 were never rated top (and always beaten by the two top ranking systems), neither for the furniture nor for the people subdomain, neither for Fluency nor Adequacy. In fact, overall, HUMAN-1 and HUMAN-2 ranked fourth and sixth (out of 8) for Adequacy, and third and fourth for Fluency.

From this it is clear that systems were able to produce better quality outputs, at least as perceived by our assessors, than the data they were trained on. This was possible, because most systems used information other than frequency of occurrence in the training data; some systems hardwired defaults such as always including the type (e.g. *chair*) of an entity in a description; and many of the low-quality aspects of descriptions in the training data, provided they were of sufficiently low frequency, would have simply 'come out in the statistical wash'. However, in computing the automatic intrinsic scores, system outputs

[19] Some of the more idiosyncratic examples from the test data include the following: "a red chair, if you sit on it, your feet would show the south east", "male dark hair grey beard and black rimmed glasses", "left hand photo on 2nd row", "the dark haired, younger looking aged gentleman with spectacles", "the picture of the person on the far right in the top row is surrounded by a red border", "top right".

are assessed in terms of their similarity with the original corpus data (which our human judges considered less than ideal, certainly worse than some systems).

If the test data reference descriptions had been such that the human assessors rated them highly (higher than the system outputs at least), then it is reasonable to expect that scores computing string similarity with those reference descriptions would have also had some correlation with the human assessors' scores. Given that there is good correlation between the assessors' scores and extrinsic scores, one might then also find some correlation between the extrinsic scores and the automatic intrinsic ones.

While the quality of the TUNA data may offer a *plausible* explanation, we have at present no way of confirming that this is the *right* explanation. The way to test it would be to redo the experiments with a test set that the human assessors judge to be of high quality. Here, however, there are two issues to be considered. First, even if such an experiment revealed a better correlation between automatic and human intrinsic scores, it may of course have no effect on the lack of correlation between automatic intrinsic and extrinsic scores. Second, it is an open question whether such a test set could be produced for a task such as REG, without introducing artificially stringent instructions to participants aiming to minimise certain types of variation (for example, instructions to the effect that only full definite NPs can be produced). Individual variation is a well-known problem in the use of corpora texts as reference outputs for NLG. This has been discussed in the context of REG by Viethen and Dale [76], and also by Reiter and Sripada [77] in the context of a different task. As we suggested above, some of the variation observed in our test data may be due to the fact that the authors represented in the corpus were self-reported native *or* fluent speakers of English, at least when the original corpus was constructed. However, it seems unlikely that this is the sole explanation, since the native speakers were in the majority, and the test sets constructed for TUNA'08 and TUNA'09 consisted exclusively of data from native speakers. Moreover, variation also seems to arise in more controlled studies of reference. For example, a recent study which replicated the TUNA experiment with Dutch speakers, under controlled laboratory conditions, found substantial variation among authors as in the original TUNA corpus [78]. It is quite possible that the kind of telegraphic style we observed among some of the authors in the TUNA corpus is simply due to the nature of the task which, unlike tasks involving the production of longer sentences or texts, can be carried out (and indeed is often carried out in normal everyday speech) using relatively 'degraded' forms without necessarily compromising the success of the main communicative goal, namely, to identify an object.

4.4 Summary and Implications

The three sets of results reported above can be summarised as follows.

The role of minimality: While minimality or brevity emerged as the main theoretically motivated measure of quality of a referring expression in the traditional REG literature, our findings suggest that it negatively impacts the humanlikeness

of a referring expression, as reflected in significant negative correlations. In spite of a long tradition of psycholinguistic research showing that human speakers tend to overspecify when referring, our results remain topical for REG, where appeals to brevity have continued to be made until recently (compare, for example, the psycholinguistically-motivated arguments given by Dale and Reiter in favour of overspecification in the Incremental Algorithm [38] with the argument for Brevity offered by Gardent [72] for the generation of plurals, where minimality is adopted in order to avoid excessive complexity in the logical forms generated by a content determination algorithm).

While brevity no doubt has *some* role to play (one would not argue, for example, that *all* possible attributes should be included in a description by default), looking at human-produced descriptions shows that they do not tend to opt for the absolute minimal number of attributes required to identify a referent. Our results furthermore give some preliminary indication that descriptions that are minimal in this sense may actually delay identification (Section 4.1).

Automatic measures of humanlikeness: Concerns with the validity of automatically computed metrics have been voiced in other fields, notably Summarisation [8] and Machine Translation [5]. Alongside several existing reports of discrepancies between intrinsic and extrinsic evaluation measures, our results raise the possibility that automatic, corpus-based metrics of humanlikeness focus on very different aspects of the quality of human referring expressions than those tapped into by extrinsic, task-based measures. Nevertheless, we have argued that the results reported here need to be interpreted with caution and require further follow-up research to avoid some potential methodological pitfalls.

Human intrinsic measures of quality: The relationship between human intrinsic measures of quality (in the form of judged appropriateness) and extrinsic measures in our 2009 results appears to be more straightforwardly interpretable. There therefore appears to be an interesting asymmetry between the two classes of intrinsic measures considered in this chapter, namely corpus-based/automatic and human. We have discussed some possible reasons for this, foremost among which is the fact that our reference material (the human-authored referring expressions in the test data part of the TUNA corpus) are not regarded as high quality by human judges. Another possible explanation is that there is an underlying asymmetry in the standards of evaluation adopted in the two classes of intrinsic measures. Corpus-based, automatic measures compare outputs to what people do (i.e. to human production), whereas in a judgement experiment, participants are asked to tap into their own individual subjective views of what makes a referring expression good or bad.

Because of the methodological caveats we have noted with some aspects of our study, a replication of the studies reported here involving a larger sample and a new dataset produced in a more controlled setting seems particularly worthwhile, given that the question of the relationship between different classes of methods is crucial to a better understanding of evaluation methodology in NLG.

5 Conclusion

This chapter has given an overview of the structure and scope of the three editions of the TUNA Shared Task Evaluation Challenges. We have argued that such exercises in comparative evaluation can be highly beneficial, not least because a concentrated effort on the part of several research teams on a single problem can enhance progress and the range of approaches to the problem can be broadened as a result. In the case of Referring Expression Generation, we believe that these events have helped to broaden the field by including realisation in addition to the more familiar attribute selection task and also through the variety of algorithmic approaches adopted by the various participants, including many trainable systems (new to REG).

These events have yielded a rich source of comparative evaluation data based on a large number of different evaluation measures. Such data can function both as a baseline against which to compare future systems, and as a repository to be exploited for further follow-up investigation. We have presented results from some of our follow-up investigations [6, 31] in this chapter, where we focused on the relationship between different kinds of evaluation criteria. Our results suggest that different classes of methods give different results whose relationship is not always predictable or transparent. This is especially true of the relationship between metrics computed against corpora and measures obtained through task-based experiments. Unlike corpus-based automatic metrics, human judgements constitute a class of intrinsic measures which do evince a systematic relationship with extrinsic ones. In future work, we are planning to follow up these studies further, using a new corpus and experimental data.

Acknowledgments

We are grateful to all the participants in the three editions of the TUNA Challenges, whose enthusiasm and input made the work reported here possible. Thanks to Ehud Reiter and Jette Viethen, who co-organised the first event, and to Irene Langkilde-Geary for her work on the template-based realiser used in TUNA-AS'07. The participants of our extrinsic evaluation experiments were drawn from the staff and students of UCL, Sussex and Brighton universities. We are also grateful to Emiel Krahmer for helpful comments on earlier drafts of this paper. Referring Expression Generation Challenges 2008 and Generation Challenges 2009 were organised with the financial support of the UK Engineering and Physical Sciences Research Council (EPSRC).

References

1. Spärck Jones, K., Galliers, J.R.: Evaluating natural language processing systems: An analysis and review. Springer, Berlin (1996)
2. Papineni, S., Roukos, T., Ward, W., Zhu., W.: BLEU: A method for automatic evaluation of machine translation. In: Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL'02). (2002) 311–318

3. Doddington, G.: Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. In: Proceedings of the 2nd International Conference on Human Language Technology Research (HLT'02). (2002) 138–145
4. Lin, C.Y., Hovy, E.: Automatic evaluation of summaries using n-gram co-occurrence statistics. in. In: Proceedings of HLT-NAACL-03. (2003) 71–78
5. Calliston-Burch, C., Osborne, M., Koehn, P.: Re-evaluating the role of BLEU in machine translation research. In: Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics (EACL-06). (2006) 249–256
6. Belz, A., Gatt, A.: Intrinsic vs. extrinsic evaluation measures for referring expression generation. In: Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics (ACL'08). (2008) 197–200
7. Reiter, E., Belz, A.: An investigation into the validity of some metrics for automatically evaluating Natural Language Generation systems. Computational Linguistics **35**(4) (2009) 529–558
8. Dorr, B.J., Monz, C., President, S., Schwartz, R., Zajic, D.: A methodology for extrinsic evaluation of text summarization: Does ROUGE correlate? In: Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarisation. (2005) 1–8
9. Belz, A.: That's nice ... what can you do with it? Computational Linguistics **35**(1) (2009) 111–118
10. Lester, J., Porter, B.: Developing and empirically evaluating robust explanation generators: The KNIGHT experiments. Computational Linguistics **23**(1) (1997) 65–101
11. Callaway, C.B., Lester, J.C.: Narrative prose generation. Artificial Intelligence **139**(2) (2002) 213–252
12. Foster, M.: Automated metrics that agree with human judgements on generated output for an embodied conversational agent. In: Proceedings of the 5th International Conference on Natural Language Generation (INLG'08). (2008) 95–103
13. Reiter, E., Sripada, S., Hunter, J., Yu, J., Davy, I.: Choosing words in computer-generated weather forecasts. Artificial Intelligence **167** (2005) 137–169
14. Belz, A., Kow, E., Viethen, J., Gatt, A.: Generating referring expressions in context: The GREC task evaluation challenges. In Krahmer, E., Theune, M., eds.: Empirical Methods in Natural Language Generation. Volume 5980 of LNCS. Springer, Berlin / Heidelberg (2010)
15. Langkilde-Geary, I.: An empirical verification of coverage and correctness for a general-purpose sentence generator. In: Proceedings of the 2nd International Conference on Natural Language Generation (INLG'02). (2002)
16. Callaway, C.B.: Evaluating coverage for large symbolic NLG grammars. In: Proceedings of the 18th International Joint Conference on Artificial Intelligence (IJCAI'03). (2003) 811–817
17. Belz, A.: Statistical generation: Three methods compared and evaluated. In: Proceedings of the 10th European Workshop on Natural Language Generation (ENLG'05). (2005) 15–23
18. Cahill, A., van Genabith, J.: Robust pcfg-based generation using automatically acquired lfg approximations. In: Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics (COLING/ACL'06). (2006) 1033–1040
19. White, M., Rajkumar, R., Martin, S.: Towards broad coverage surface realization with CCG. In: Proceedings of the Workshop on Using Corpora for NLG: Language Generation and Machine Translation (UCNLG+MT). (2007)

20. Jordan, P., Walker, M.: Learning attribute selections for non-pronominal expressions. In: Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics. (2000)
21. Gupta, S., Stent, A.J.: Automatic evaluation of referring expression generation using corpora. In: Proceedings of the 1st Workshop on Using Corpora in NLG (UCNLG'05). (2005) 1–6
22. Viethen, J., Dale, R.: Algorithms for generating referring expressions: Do they do what people do? In: Proceedings of the 4th International Conference on Natural Language Generation (INLG-06). (2006) 63–70
23. Gatt, A., van der Sluis, I., van Deemter, K.: Evaluating algorithms for the generation of referring expressions using a balanced corpus. In: Proceedings of the 11th European Workshop on Natural Language Generation (ENLG'07). (2007) 45–56
24. Reiter, E., Robertson, R., Osman, L.: Lessons from a failure: Generating tailored smoking cessation letters. Artificial Intelligence **144** (2003) 41–58
25. Portet, F., Reiter, E., Gatt, A., Hunter, J., Sripada, S., Freer, Y., Sykes, C.: Automatic generation of textual summaries from neonatal intensive care data. Artificial Intelligence **173**(7–8) (2009) 789–816
26. Karasimos, A., Isard, A.: Multi-lingual evaluation of a natural language generation system. In: Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC'04). (2004)
27. Stock, O., Zancanaro, M., Busetta, P., Callaway, C., Krueger, A., Kruppa, M., Kuflik, T., Not, E., Rocchi, C.: Adaptive, intelligent presentation of information for the museum visitor in PEACH. User Modeling and User-Adapted Interaction **17**(3) (2007) 257–304
28. Belz, A., Reiter, E.: Comparing automatic and human evaluation of NLG systems. In: Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics (EACL'06). (2006) 313–320
29. Belz, A., Kow, E.: System-building cost vs. output quality in data-to-text generation. In: Proceedings of the 12th European Workshop on Natural Language Generation (ENLG'09). (2009) 16–24
30. Belz, A., Kilgarriff, A.: Shared task evaluations in HLT: Lessons for NLG. In: Proceedings of INLG'06. (2006) 133–135
31. Gatt, A., Belz, A.: Attribute selection for referring expression generation: New algorithms and evaluation methods. In: Proceedings of the 5th International Conference on Natural Language Generation (INLG'08). (2008) 50–58
32. Dale, R., White, M., eds.: Shared Tasks and Comparative Evaluation in Natural Language Generation: Workshop Report. (2007)
33. Koller, A., Striegnitz, K., Byron, D., Cassell, J., Dale, R., Moore, J., Oberlander, J.: The first challenge on generating instructions in virtual environments. In Krahmer, E., Theune, M., eds.: Empirical Methods in Natural Language Generation. Volume 5980 of LNCS. Springer, Berlin / Heidelberg (2010)
34. Appelt, D.: Planning English referring expressions. Artificial Intelligence **26**(1) (1985) 1–33
35. Appelt, D., Kronfeld, A.: A computational model of referring. In: Proceedings of the 10th International Joint Conference on Artificial Intelligence (IJCAI-87). (1987) 640–647
36. Kronfeld, A.: Conversationally relevant descriptions. In: Proceedings of the 27th Annual Meeting of the Association for Computational Linguistics (ACL'89). (1989) 60–67
37. Dale, R.: Cooking up referring expressions. In: Proceedings of the 27th Annual Meeting of the Association for Computational Linguistics, ACL-89. (1989) 68–75

38. Dale, R., Reiter, E.: Computational interpretation of the Gricean maxims in the generation of referring expressions. Cognitive Science **19**(8) (1995) 233–263
39. Bohnet, B., Dale, R.: Viewing referring expression generation as search. In: Proceedings of the 19th International Joint Conference on Artificial Intelligence (IJCAI'05). (2005) 1004–1009
40. Jordan, P.W., Walker, M.: Learning content selection rules for generating object descriptions in dialogue. Journal of Artificial Intelligence Research **24** (2005) 157–194
41. van der Sluis, I., Gatt, A., van Deemter, K.: Evaluating algorithms for the generation of referring expressions: Going beyond toy domains. In: Proceedings of the Conference on Recent Advances in Natural Language Processing (RANLP'07). (2007)
42. Jordan, P.W.: Can nominal expressions achieve multiple goals? In: Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics (ACL'00). (2000) 142–149
43. Reips, U.D.: The Web Experimental Psychology Lab: Five years of data collection on the Internet. Behavioral Research Methods and Computers **33**(2) (2001) 201–211
44. van der Sluis, I., Krahmer, E.: Generating multimodal referring expressions. Discourse Processes. **44**(3) (2007) 145–174
45. von Stutterheim, C., Mangold-Allwinn, R., Barattelli, S., Kohlmann, U., Kölbing, H.G.: Reference to objects in text production. Belgian Journal of Linguistics **8** (1993) 99–125
46. Maes, A., Arts, A., Noordman, L.: Reference management in instructive discourse. Discourse Processes **37**(2) (2004) 117–144
47. Pechmann, T.: Incremental speech production and referential overspecification. Linguistics **27** (1989) 89–110
48. Belz, A., Gatt, A.: The attribute selection for GRE challenge: Overview and evaluation results. In: Proceedings of UCNLG+MT: Language Generation and Machine Translation. (2007) 75–83
49. Gatt, A., Belz, A., Kow, E.: The TUNA Challenge 2008: Overview and evaluation results. In: Proceedings of the 5th International Conference on Natural Language Generation (INLG'08). (2008) 198–206
50. Gatt, A., Belz, A., Kow, E.: The TUNA-REG Challenge 2009: Overview and evaluation results. In: Proceedings of the 12th European Workshop on Natural Language Generation (ENLG'09). (2009) 198–206
51. van Deemter, K., Gatt, A.: Content determination in GRE: Evaluating the evaluator. In: Proceedings of the 2nd UCNLG Workshop: Language Generation and Machine Translation. (2007)
52. Pereira, D.B., Paraboni, I.: From TUNA attribute sets to Portuguese text: A first report. In: Proceedings of the 5th International Conference on Natural Language Generation (INLG'08). (2008) 232–234
53. Krahmer, E., van Erk, S., Verleg, A.: Graph-based generation of referring expressions. Computational Linguistics **29**(1) (2003) 53–72
54. King, J.: OSU-GP: Attribute selection using genetic programming. In: Proceedings of the 5th International Conference on Natural Language Generation (INLG'08). (2008) 225–226
55. Hervás, R., Gervás, P.: Evolutionary and case-based approaches to REG. In: Proceedings of the 12th European Workshop on Natural Language Generation (ENLG'09). (2009) 187–188

56. Spanger, P., Kurosawa, T., Tokunaga, T.: TITCH: Attribute selection based on discrimination power and frequency. In: Proceedings of UCNLG+MT: Language Generation and Machine Translation. (2008) 98–100
57. de Lucena, D., Paraboni, I.: USP-EACH frequency-based greedy attribute selection for referring expressions generation. In: Proceedings of the 5th International Conference on Natural Language Generation (INLG'08). (2008) 219–220
58. Theune, M., Touset, P., Viethen, J., Krahmer, E.: Cost-based attribute selection for generating referring expressions (GRAPH-FP and GRAPH-SC). In: Proceedings of UCNLG+MT: Language Generation and Machine Translation. (2007) 95–97
59. Bohnet, B.: IS-FBN, IS-FBS, IS-IAC: The adaptation of two classic algorithms for the generation of referring expressions in order to produce expressions like humans do. In: Proceedings of UCNLG+MT: Language Generation and Machine Translation. (2007) 84–86
60. Bohnet, B.: The fingerprint of human referring expressions and their surface realization with graph transducers. In: Proceedings of the 5th International Conference on Natural Language Generation (INLG'08). (2008) 207–210
61. Fabbrizio, G.D., Stent, A.J., Bangalore, S.: Referring expression generation using speaker-based attribute selection and trainable realization (ATT-REG). In: Proceedings of the 5th International Conference on Natural Language Generation (INLG'08). (2008) 211–214
62. Fabbrizio, G.D., Stent, A.J., Bangalore, S.: Trainable speaker-based referring expression generation. In: Proceedings of the 12th Conference on Computational Natural Language Learning (CONLL'08). (2008) 151–158
63. Kelleher, J., Namee, B.M.: Referring expression generation challenge 2008 DIT system descriptions. In: Proceedings of the 5th International Conference on Natural Langauge Generation (INLG'08). (2008) 221–224
64. Passonneau, R.: Measuring agreement on set-valued items (MASI) for semantic and pragmatic annotation. In: Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC'06). (2006)
65. Grice, H.: Logic and conversation. In Cole, P., Morgan, J., eds.: Syntax and Semantics: Speech Acts. Volume III. Academic Press (1975)
66. Bard, E.G., Robertson, D., Sorace, A.: Magnitude estimation of linguistic acceptability. Language **72**(1) (1996) 32–68
67. Gatt, A., van Deemter, K.: Lexical choice and conceptual perspective in the generation of plural referring expressions. Journal of Logic, Language and Information **16**(4) (2007) 423–443
68. Tanenhaus, M.K., Spivey-Knowlton, M.J., Eberhard, K.M., Sedivy, J.G.: Integration of visual and linguistic information in spoken language comprehension. Science **268** (1995) 1632–1634
69. Black, A., Taylor, P., Caley, R.: The Festival speech synthesis system: System documentation. Technical Report 1.4 edition., University of Edinburgh (1999)
70. Belke, E., Meyer, A.: Tracking the time course of multidimensional stimulus discrimination: Analysis of viewing patterns and processing times during same-different decisions. European Journal of Cognitive Psychology **14**(2) (2002) 237–266
71. Engelhardt, P.E., Bailey, K., Ferreira, F.: Do speakers and listeners observe the Gricean Maxim of Quantity? Journal of Memory and Language **54** (2006) 554–573
72. Gardent, C.: Generating minimal definite descriptions. In: Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL'02). (2002) 96–103

73. van Deemter, K.: Generating referring expressions: Boolean extensions of the incremental algorithm. Computational Linguistics **28**(1) (2002) 37–52
74. Miyao, Y., Saetre, R., Sagae, K., Matsuzaki, T., Tsujii, J.: Task-oriented evaluations of syntactic parsers and their representations. In: Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics (ACL'08). (2008) 46–54
75. Law, A.S., Freer, Y., Hunter, J., Logie, R., McIntosh, N., Quinn, J.: A comparison of graphical and textual presentations of time series data to support medical decision making in the neonatal intensive care unit. Journal of Clinical Monitoring and Computing **19** (2005) 183–194
76. Viethen, J., Dale, R.: Evaluation in natural language generation: Lessons from referring expression generation. Traitement Automatique des Langues **48**(1) (2007) 141–160
77. Reiter, E., Sripada, S.: Should corpora texts be gold standards for NLG? In: Proceedings of the 2nd International Conference on Natural Language Generation (INLG'02). (2002)
78. Koolen, R., Gatt, A., Goudbeek, M., Krahmer, E.: Need I say more? On factors causing referential overspecification. In: Proceedings of the Workshop on Production of Referring Expressions: Bridging Computational and Psycholinguistic Approaches (PRE-COGSCI'09). (2009)