Generating Referring Expressions in Context: The GREC Task Evaluation Challenges

Anja Belz,[1] Eric Kow,[1] Jette Viethen[2] and Albert Gatt[3,4]

[1] University of Brighton, Brighton BN2 4GJ, UK
{a.s.belz,e.y.kow}@brighton.ac.uk
[2] Macquarie University, Sydney NSW 2109, Australia jviethen@ics.mq.edu.au
[3] Institute of Linguistics, Centre for Communication Technology, University of Malta
albert.gatt@um.edu.mt
[4] Communication and Cognition, Faculty of Arts, Tilburg University, Netherlands

Abstract. Until recently, referring expression generation (REG) research focused on the task of selecting the semantic content of definite mentions of listener-familiar discourse entities. In the GREC research programme we have been interested in a version of the REG problem definition that is (i) grounded within discourse context, (ii) embedded within an application context, and (iii) informed by naturally occurring data. This paper provides an overview of our aims and motivations in this research programme, the data resources we have built, and the first three shared-task challenges, GREC-MSR'08, GREC-MSR'09 and GREC-NEG'09, we have run based on the data.

1 Background

Referring Expression Generation (REG) is one of the most lively and thriving subfields of Natural Language Generation (NLG). Traditionally, it has addressed the following question:

> [G]iven a symbol corresponding to an intended referent, how do we work out the semantic content of a referring expression that uniquely identifies the entity in question? [5, p. 1004]

Realisation, i.e. turning the resulting semantic content representation into a string of words, is not part of this problem specification, and REG has moreover predominantly considered the task in isolation—taking into account neither the discourse context (it was assumed that attributes were being selected for definite mentions of listener-familiar entities), nor the context of the language generation process (it was assumed that a REG module is called at some point in the generation process and that referent, potential distractors, and their possible attributes will be provided as parameters to the REG module).

In the 1990s, REG research looked at two main factors in selecting attributes (semantic content): unique identification (of the intended referent from a set including possible distractors), and brevity [9, 31]. The most influential of these algorithms, the Incremental Algorithm (IA) [10], originally just selected attributes

for a single entity from a given set, but a range of extensions have been reported, including van Deemter's SET algorithm which can generate REs to sets of entities [11], and Siddharthan and Copestake's algorithm [32] which is able to identify attributes that are particularly discriminating given the entities in the contrast set of distractor entities.

Work in the 2000s increasingly took into account that there is more to REG than attribute selection, identification and brevity. Krahmer and Theune [21] moved away from the simplifying assumption, made by Dale and Reiter among others, that the contextually specified set of salient destractors would be provided to the REG algorithm. Their context-sensitive version of the IA took context into account, replacing the requirement that the intended referent be the *only* entity that matches the RE, to the requirement that it be the *most salient* in a given context. Jordan [18] showed that REs used by people do not always follow the brevity principle: she found a large proportion of over-specified redescriptions in the Coconut corpus of dialogues and showed that some dialogue states and communicative goals make over-specified REs more likely. Viethen & Dale [37] pointed out that the question why people choose different REs in different contexts has not really been addressed:

> Not only do different people use different referring expressions for the same object, but the same person may use different expressions for the same object on different occasions. Although this may seem like a rather unsurprising observation, it has never, as far as we are aware, been taken into account in the development of any algorithm for generation of referring expressions. [37, p. 119]

Researchers also started looking at real REG data, e.g. Viethen & Dale [37] and Gatt et al. [12] collected corpora of referring expressions elicited by asking participants to describe entities in scenarios of furniture items and faces. Others have looked at REs in discourse context. Nenkova's thesis [26] looked at rewriting mentions of people in extractive summaries with the aim of improving them within their new context, focusing on first mentions. Belz & Varges [3] collected a corpus of Wikipedia articles in order to investigate the question of how writers select mentions of named entities (cities, countries, rivers, people, mountains) in discourse context.

Other resources exist within which REs have been annotated in some way. In the GNOME Corpus [28, 29] different types of discourse and semantic information are annotated, including reference and semantic attributes. The corpus annotation was, for example, used to train a decision tree learner for NP modifier generation [7]. The RE annotations in the Coconut corpus represent information at the discourse level (reference and attributes used) and at the utterance level (information about dialogue state); the 400 REs with their annotations in the corpus were used to train a REG module [19]. Gupta and Stent [16] annotated both the Maptask and Coconut corpora with POS-tags, NP boundaries, referents and knowledge representations for each speaker which included values for different attributes for potential referents.

2 Overview of GREC Research Programme

Extending the existing body of REG work that has started taking discourse context and real data into account, and building on earlier work [3], in the GREC research programme (Generating Referring Expressions in Context) we have been interested in a version of the REG problem that is (i) grounded within discourse context, (ii) embedded within an application context, and (iii) informed by naturally occurring data. This paper provides an overview of our aims and motivations in this research programme, the data resources we have built (the GREC-2.0 corpus and the GREC-People corpus), and the first three shared-task challenges we have run based on these data resources (GREC-MSR'08, GREC-MSR'09 and GREC-NEG'09.

In the next two sections (Sections 3 and 4), we describe the two data resources and annotation schemes we have created for GREC. In Section 5 we outline the two GREC task definitions (GREC-MSR and GREC-NEG), and in Section 6 the evaluation procedures we have applied to the two tasks. In Section 7 we provide a brief overview of systems and results in the two GREC-MSR challenges we ran in 2008 and 2009, and in Section 8 of the GREC-NEG challenge which ran for the first time in 2009 and for the second time in 2010. In Section 9 we discuss some of the issues and outcomes of the GREC evaluations.

In the remainder of this section, we briefly introduce the common features of the two GREC tasks and summarise the differences between the annotation schemes we have developed for them.

2.1 The GREC Task in general terms

In general terms, the GREC tasks are about how to generate appropriate references to an entity in the context of a piece of discourse longer than a sentence. Rather than requiring participants to generate referring expressions from scratch, the GREC-MSR and GREC-NEG tasks provide sets of possible referring expressions for selection. Figure 1 shows a human-readable version of the GREC task: all references to *Isaac Newton* have been deleted and lists of possible referring expressions are provided instead. The task is to select a sequence of referring expressions to insert into the gaps such that the resulting text is fluent and coherent.

The immediate motivating application context for the GREC Tasks is the improvement of referential clarity and coherence in extractive summaries and multiply edited texts (such as Wikipedia articles) by regenerating referring expressions contained in them. The motivating theoretical interest for the GREC Tasks is to discover what kind of information is useful for making choices between different kinds of referring expressions in context.

In both the GREC-2.0 and the GREC-People annotation schemes, a distinction is made between *reference* and *referential expression*. A reference is 'an instance of referring' which is unique, whereas a referential expression is a word string and each reference can be realised by many different referential expressions. In the GREC corpora, each time an entity is referred to, there is a single reference,

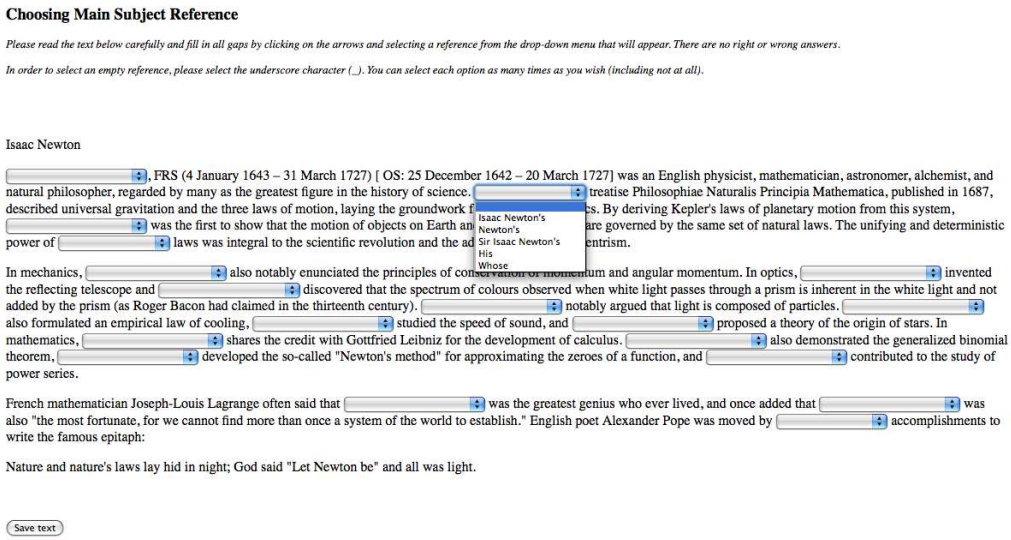

Fig. 1. Screenshot of experiment in which participants performed the GREC tasks manually.

but there may be several referring expressions corresponding to it: in the training/development data, there is a single RE for each reference (the one found in the corpus), and a set of possible alternative REs is also provided; in the test set, there are four REs for each reference (the one from the corpus and three additional ones selected by subjects in a manual selection experiment), as well as the list of alternative REs.

2.2 Summary of differences between the two GREC datasets (GREC-People and GREC-2.0)

The GREC-MSR and GREC-NEG tasks used different datasets, namely the GREC-2.0 corpus and the GREC-People corpus, respectively. The main difference between these is that in GREC-2.0 only references to the main subject of the text (MSRs) have been annotated, whereas in GREC-People, references to more than one discourse entity have been annotated. Other differences are that in GREC-People, (i) we have corrected spelling errors, (ii) the annotations have been extended to plural references, attributive complements, appositive supplements and subjects of gerund-participials, (iii) integrated dependents are included within the annotation of an RE, and (iv) the `SYNCAT` attribute has been split into `SYNCAT` and `SYNFUNC` attributes, indicating (just) syntactic category of the RE and syntactic function, respectively. See below for explanations of all these terms.

3 The GREC-2.0 Corpus

The GREC-2.0 corpus consists of 1,941 introduction sections from Wikipedia articles in five different domains (cities, countries, rivers, people and mountains). The corpus texts have been annotated for three broad categories of references to the main subject[5] of each text, called Main Subject References (MSRs). These are categories which were relatively simple to identify and achieve high inter-annotator agreement on (complete agreement among four annotators in 86% of MSRs).

The GREC-2.0 corpus has been divided into training, development and test data for the purposes of the GREC-MSR Task. The number of texts in the three data sets and the five subdomains is as follows:

	All	Mountains	People	Countries	Cities	Rivers
Total	1941	932	442	251	243	73
Training	1658	791	373	216	213	65
Development	97	46	24	12	11	4
Test	183	92	45	23	19	4

3.1 Types of referential expressions annotated

In terminology and view of grammar our approach to annotating REs relies heavily on Huddleston and Pullum's Cambridge Grammar of the English Language [17]. We have annotated three broad categories of referential expression (RE) in the GREC-2.0 corpus: (i) subject NPs, (ii) object NPs and (iii) genitive NPs including pronouns which function as subject-determiners within their matrix NP.

I Subject NPs: referring subject NPs, including pronouns and special cases of VP coordination where the same RE functions as the subject of the coordinated VPs (see Section 3.2), e.g:

1. *He was proclaimed dictator for life.*
2. *Alexander Graham Bell (March 3, 1847 - August 2, 1922) was a Scottish scientist and inventor who emigrated to Canada.*
3. *Most Indian and Bangladeshi rivers bear female names, but this one has a rare male name.*
4. *"The Eagle" was born in Carman, Manitoba and __ grew up playing hockey.*

II Object NPs: referring NPs that function as direct or indirect objects of VPs and prepositional phrases; e.g.:

1. *These sediments later deposit in the slower lower reaches of the river.*
2. *People from the city of São Paulo are called paulistanos.*
3. *His biological finds led him to study the transmutation of species.*

[5] An example of a main subject of a text in the cities domain is e.g. *London*.

III Subject-determiner genitives: genitive NPs that function as subject-determiners[6] including genitive forms of pronouns. Note that this excludes genitives that are the subject of a gerund-participial[7]:

1. *Its estimated length is 4,909 km.*
2. *The country's culture, heavily influenced by neighbours, is based on a unique form of Buddhism intertwined with local elements.*
3. *Vatican City is a landlocked sovereign city-state whose territory consists of a walled enclave within the city of Rome.*

3.2 Comments on some aspects of annotation

Some types of relative pronoun, those in supplementary relative clauses (as opposed to integrated relative clauses, see Huddleston and Pullum, 2002, p. 1058), are interpreted as anaphorically referential (I(2) and III(3) above). These differ from integrated relative clauses in that in supplementary relative clauses, the relative clause can be dropped without affecting the meaning of the clause containing it. From the point of view of generation, the meaning could be equally expressed in two independent sentences or in two clauses of which one is a supplementary relative clause. An example of the single-sentence construction is shown in (1) below, with the semantically equivalent two-sentence alternative shown in (2):

(1) *Hristo Stoichkov is a football manager and former striker who was a member of the Bulgaria national team that finished fourth at the 1994 FIFA World Cup.*
(2) *Hristo Stoichkov is a football manager and former striker. He was a member of the Bulgaria national team that finished fourth at the 1994 FIFA World Cup.*

The GREC-2.0 annotation scheme also includes 'non-realised' subject REs in a restricted set of cases of VP coordination where an RE is the subject of the coordinated VPs. Consider the following example, where the subclausal coordination in (3) is semantically equivalent to the clausal coordination in (4):

(3) *He stated the first version of the Law of conservation of mass, __ introduced the Metric system, and __ helped to reform chemical nomenclature.*
(4) *He stated the first version of the Law of conservation of mass, he introduced the Metric system, and he helped to reform chemical nomenclature.*

According to Huddleston and Pullum, utterances as in (3) can be thought of as a reduction of longer forms as in (4), even though the former are not syntactically derived by ellipsis from the latter p. 1280, and from the point of view of language analysis there is no need for an analysis involving a null anaphoric reference. The motivation for annotating the approximate place where the subject NP would be if it were realised (the gap-like underscores above) is that from a generation

[6] I.e. "they combine the function of determiner, marking the NP as definite, with that of complement (more specifically subject)." (Huddleston and Pullum (2002), p. 56)

[7] E.g. *His early career was marred by *his being involved in a variety of social and revolutionary causes.*

```
<?xml version="1.0" encoding="utf-8"?>
<!DOCTYPE TEXT SYSTEM "reg08-grec.dtd">
<TEXT ID="36">

<TITLE>Jean Baudrillard</TITLE>

<PARAGRAPH>
 <REF ID="36.1" SEMCAT="person" SYNCAT="np-subj">
  <REFEX REG08-TYPE="name" EMPHATIC="no" HEAD="nominal" CASE="plain">Jean Baudrillard</REFEX>
  <ALT-REFEX>
   <REFEX REG08-TYPE="name" EMPHATIC="no" HEAD="nominal" CASE="plain">Jean Baudrillard</REFEX>
   <REFEX REG08-TYPE="name" EMPHATIC="yes" HEAD="nominal" CASE="plain">Jean Baudrillard himself</REFEX>
   <REFEX REG08-TYPE="empty">_</REFEX>
   <REFEX REG08-TYPE="pronoun" EMPHATIC="no" HEAD="pronoun" CASE="nominative">he</REFEX>
   <REFEX REG08-TYPE="pronoun" EMPHATIC="yes" HEAD="pronoun" CASE="nominative">he himself</REFEX>
   <REFEX REG08-TYPE="pronoun" EMPHATIC="no" HEAD="rel-pron" CASE="nominative">who</REFEX>
   <REFEX REG08-TYPE="pronoun" EMPHATIC="yes" HEAD="rel-pron" CASE="nominative">who himself</REFEX>
  </ALT-REFEX>
 </REF>
 (born June 20, 1929) is a cultural theorist, philosopher, political commentator,
 sociologist, and photographer.
 <REF ID="36.2" SEMCAT="person" SYNCAT="subj-det">
  <REFEX REG08-TYPE="pronoun" EMPHATIC="no" HEAD="pronoun" CASE="genitive">His</REFEX>
  <ALT-REFEX>
   <REFEX REG08-TYPE="name" EMPHATIC="no" HEAD="nominal" CASE="genitive">Jean Baudrillard's</REFEX>
   <REFEX REG08-TYPE="pronoun" EMPHATIC="no" HEAD="pronoun" CASE="genitive">his</REFEX>
   <REFEX REG08-TYPE="pronoun" EMPHATIC="no" HEAD="rel-pron" CASE="genitive">whose</REFEX>
  </ALT-REFEX>
 </REF>
 work is frequently associated with postmodernism and post-structuralism.
</PARAGRAPH>

</TEXT>
```

Fig. 2. Example text from GREC-2.0 corpus.

perspective there is a choice to be made about whether to realise the subject NP in the second (and subsequent) coordinate(s) or not. Note that only cases of subclausal coordination at the level of VPs have been annotated in this way. Therefore these are all cases where *only* the subject NP is 'missing'.[8]

There are some items that could be construed as main subject reference which we decided not to include in teh GREC-2.0 annotation scheme. These include those that are, according to Huddleston and Pullum, true gaps and ellipses, adjective and noun modifiers, and implicit or anaphorically derivable references (other than those mentioned above). Some of these we added to the annotation in the separate GREC-People corpus (see Section 4 below). Furthermore, we did not annotate the following types of text elements at all: the main title of the article, titles of books, films, etc. mentioned in the article; citations from articles, books, films, etc.; names and titles of organisations and persons (except where they are used in their entirety to refer to the main subject).

3.3 XML format

The XML format described below was intended only for the purpose of the GREC-MSR Task. While attempts have been made to make it linguistically plausible and generic, certain aspects of it have been determined solely by the requirements of the GREC-MSR Task.

Figure 2 shows one of the texts from the GREC-2.0 corpus. Each item in the corpus is an XML annotated text file, and is of document type `TEXT`.

A `TEXT` is composed of one `TITLE` followed by any number of `PARAGRAPH`s. A `TITLE` is just a string of characters. A `PARAGRAPH` is any combination of character strings and `REF` elements. The `REF` element indicates a reference, in the sense of 'an instance of referring' (as discussed above). A `REF` is composed of one `REFEX` element (the 'selected' referential expression for the given reference; in the training data texts it is just the referential expression found in the corpus) and one `ALT-REFEX` element which in turn is a list of `REFEX`s (alternative referential expressions obtained by other means, as explained in Section 4.3).

The attributes of the `REF` element are `ID`, a unique reference identifier taking integer values; `SEMCAT`, indicating the semantic category of the referent and ranging over `city`, `country`, `river`, `person`, and `mountain`; and `SYNCAT`, the syntactic category required of referential expressions for the referent in this context (values `np-obj, np-subj, subj-det`[9]).

The `SYNCAT` attribute does not so much indicate a property of reference, as a constraint on the referential expressions that can realise it in a given context. Because in the GREC-MSR Task the context is fully realised text, the constraint is on syntactic category.

The distinction between reference and referential expression is useful, because a single reference can have multiple possible realisations, but also because the two have distinct properties. For example, a reference does not have lexical and syntactic properties, whereas referential expressions do; a distinction between reference and referential expression in the annotation scheme allows such properties to be annotated only where appropriate.

A `REFEX` element indicates a referential expression (a word string that can be used to refer to an entity). It has four attributes. `HEAD` is the category of the head of the RE (values: `nominal`, `pronoun`, `rel-pron`). `CASE` indicates the case of the head (values for pronouns: `nominative`, `accusative`, `genitive`; for nominals: `plain`, `genitive`). `EMPHATIC` is a Boolean attribute and indicates whether the RE is emphatic. In the GREC-2.0 corpus, the only type of RE that has this attribute is one which incorporates a reflexive pronoun used emphatically (e.g. *India itself*).

The `REGO8-TYPE` attribute (values `name`, `common`, `pronoun`, `empty`) indicates basic RE type as required for the GREC-MSR task definition. The choice of types is motivated by the hypothesis that one of the most basic decisions to be taken in

[8] E.g. we would not annotate a non-realised RE in *She wrote books for children and books for adults.*

[9] These stand for NP in object position, NP in subject position and NP that is both a determiner and a subject, respectively. See Section 3.1 for explanation of these terms.

RE selection for named entities is whether to use an RE that includes a name, such as *Modern India* (the corresponding `REG08-TYPE` value is `name`); whether to go for a common-noun RE, i.e. with a category noun like *country* as the head (`common`); whether to pronominalise the RE (`pronoun`); or whether it can be left unrealised (`empty`).

Finally, an `ALT-REFEX` element is a list of `REFEX` elements (corresponding to different possible realisations).

4 The GREC-People Corpus

The GREC-People corpus (a separate corpus that has no overlap with GREC-2.0) consists of 1,000 annotated introduction sections from Wikipedia articles in the category People. Each text therefore has a person as the main subject. There are three subcategories: inventors, chefs and early music composers. For the purposes of the GREC-NEG competitions, the GREC-People corpus was divided into training, development and test data. The number of texts in the three data sets and the three subdomains are as follows:

	All	Inventors	Chefs	Composers
Total	1,000	307	306	387
Training	809	249	248	312
Development	91	28	28	35
Test	100	31	30	39

As in GREC-2.0, we have annotated mentions of people by marking up the word strings that function as referential expressions (REs) and annotating them with coreference information as well as syntactic and semantic features. Since the subject of each text is a person, there is at least one coreference chain in each text. The numbers of coreference chains (entities) in the 900 texts in the training and development sets are as follows (e.g. there are 38 texts with 5 person discourse entities):

x coref chains	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23
in y texts	437	192	80	63	38	31	16	18	4	7	9	1	1	0	0	0	0	0	0	1	1	0	1

The texts vary greatly in length, from 13 words to 935, the average being 128.98 words.

4.1 Annotation of Referring Expressions in GREC-People

This section describes the different types of referring expression (RE) that we annotated in the GREC-People corpus. As in GREC-2.0, we relied on Huddleston and Pullum's work for terminology and view of syntax. The manual annotations were automatically checked and converted to the XML format described in Section 4.3 (which encodes slightly less information, as explained below).

In the example sentences in the following sections, (unbroken) underlines are used for REs that are an example of the specific type of RE they are intended to illustrate, whereas dashed underlines are used for other REs that are also annotated in the corpus. Coreference between REs is indicated by subscripts $i, j, ...$ immediately to the right of an underline (the scope of the coindexing variables is one sentence, i.e. an i in one example sentence does not represent the same entity as an i in another example sentence). Square brackets indicate supplements. The syntactic component relativised by a relative pronoun is indicated by vertical bars. Supplements and their anchors (in the case of appositive supplements), and relative clauses and the component they relativise (in the case of relative-clause supplements) are co-indexed by superscript $x, y, ...$ Dependents integrated in an RE are indicated by curly brackets. Both supplements and dependents are highlighted in boldface font where they specifically are being discussed. All terms are explained below.

In the XML format of the annotations, the beginning and end of a reference is indicated by `<REF><REFEX>...` `</REFEX></REF>` tags, and other properties discussed in the following sections (such as syntactic category etc.) are encoded as attributes on these tags (for full details see Section 4.3 below). For the GREC-NEG'09 Task we decided not to transfer the annotations of integrated dependents and relative clauses to the XML format. Such dependents are included within `<REFEX>...</REFEX>` annotations where appropriate, but without being marked up as separate constituents.

We distinguish the following syntactic categories and functions:

I Subject NPs: referring subject NPs, including pronouns and special cases of VP coordination where the same referring expression functions as the subject of the coordinated VPs. For example:

1. *He$_i$ was born in Ramsay township, near Almonte, Ontario, Canada, the eldest son of $|$Scottish immigrants, {John Naismith and Margaret Young}$|^x_{j,k}$ [who$_{j,k}$ had arrived in the area in 1851 and __$_{j,k}$ worked in the mining industry]^x.*
2. *The Banū Mūsā brothers$_{i,j,k}$ were three 9th century Persian scholars, of Baghdad, active in the House of Wisdom.*

Ia Subjects of gerund-participials:

1. *His$_i$ research on hearing and speech eventually culminated in Bell$_i$ being awarded the first U.S. patent for the invention of the telephone in 1876.*
2. *Fessenden$_i$ used the alternator-transmitter to send out a short program from Brant Rock, which included his$_i$ playing the song O Holy Night on the violin ... .*
3. *Many of his$_i$ scientific contemporaries disliked him$_i$, due in part to his$_i$ using the title Professor which technically he$_i$ wasn't entitled to do.*

II Object NPs: referring NPs including pronouns that function as direct or indirect objects of VPs and prepositional phrases; e.g.:

1. *He$_i$ entrusted them$_{j,k,l}$ to Ishaq bin Ibrahim al-Mus'abi^x_m, [a former governor of Baghdad]^x_m.*

2. *He$_i$ was the son of |Nasiruddin Humayun|^x_j [whom$_j$ he$_j$ succeeded as ruler of the Mughal Empire from 1556 to 1605]^x.*

IIa Reflexive pronouns:

1. *He$_i$ committed himself$_i$ to design and development of rocket systems.*
2. *Smith$_i$ called himself$_i$ the "Komikal Konjurer".*

III Subject-determiner genitives:

1. *They$_{i,j,k}$ shared the 1956 Nobel Prize in Physics for their$_{i,j,k}$ invention.*
2. *He$_i$ is best known as |a pioneer of human-computer interaction|^x [whose$_i$ team developed hypertext, networked computers, and precursors to GUIs]^x.*
3. *On the eve of his$_i$ death in 1605, the Mughal empire spanned almost 500 million acres (doubling during Akbar's$_i$ reign).*

Note that this category excludes cases where the term has become lexicalised, such as *the so-called "Newton's method"; Koch's postulates*, which we take not to contain an embedded reference to a person.

IIIa REs in composite nominals: this is the only type of RE we have annotated that is not an NP, but a nominal. This type functions as integrated attributive complement, e.g.:

1. *The company was sold to Westinghouse in 1920, and the next year its assets, including numerous important Fessenden$_i$ patents, were sold to the Radio Corporation of America, which also inherited the Fessenden$_i$ legal proceedings.*
2. *The Eichengrün$_i$ version was ignored by historians and chemists until 1999.*
3. *These flights demonstrated the controllability of the Montgomery$_i$ design ... .*

Note that this category excludes cases where the term has become lexicalised: *the Nobel Prizes; the Gatling gun; the Moog synthesizer.*

In contrast to GREC-2.0 we also annotated supplements and RE-internal dependents, as described in detail in the GREC-NEG'09 participants' pack.[10]

4.2 Further explanation of some aspects of the annotations

Nested references: As can be seen from some of the previous examples, we annotated all embedded references, e.g.:

1. *He$_i$ was named after his$_i$ maternal grandfather^x_j [Shaikh Ali Akbar Jami]^x_j.*
2. *born in The Hague as the son of Constantijn Huygens^x_i, [a friend of René Descartes$_j$]^x_i*
3. *after European pioneers such as George Cayley's$_i$ coachman$_j$*

The maximum depth of embedding in the GREC-People corpus is 3.

Plural REs: We annotated all plural REs that refer to groups of people where the number of group members is known.

[10] The complete Participants' Packs can be downloaded here: `http://www.itri.brighton.ac.uk/home/Anja.Belz`.

Unnamed references and indefinites: We annotated all mentions of individual person entities even if they are not actually named anywhere in the text, including cases of both definite and indefinite references.

1. *The resolution's sponsor*$_i$ *described it as ...*
2. *On 25 December 1990 he*$_i$ *implemented the first successful communication between an HTTP client and server via the Internet with the help of Robert Cailliau*$_j$ *and a* $\{$*young*$\}$ *student staff* $\{$*at CERN*$\}_k$.

4.3 XML Annotation

Each item in the corpus is an XML annotated text file (an example is shown in Figure 3), which is of type `GREC-ITEM`. A `GREC-ITEM` consists of a `TEXT` element followed by an `ALT-REFEX` element. A `TEXT` has one attribute (an `ID` unique within the corpus), and is composed of one `TITLE` followed by any number of `PARAGRAPH`s. A `TITLE` is just a string of characters. A `PARAGRAPH` is any combination of character strings and `REF` elements.

The `REF` element is composed of one `REFEX` element. The attributes of the `REF` element are shown in Figure 4. `ENTITY` and `MENTION` together constitute a unique identifier for a reference within a text; together with the `TEXT ID`, they constitute a unique identifier for a reference within the entire corpus.

A `REFEX` element indicates a referential expression (a word string that can be used to refer to an entity), and has two attributes: `REG08-TYPE` is as defined for GREC-MSR (see Section 3); `CASE` indicates the case of the head (values for pronouns: `nominative`, `accusative`, `genitive`; for nominals: `plain`, `genitive`; for any 'empty' reference: `nocase`).

We allow arbitrary-depth embedding of references. This means that a `REFEX` element may have `REF` element(s) embedded in it. See below on embedding in `REFEX` elements contained in `ALT-REFEX` lists.

An `ALT-REFEX` element is a list of `REFEX` elements. For the GREC-NEG Task, these are obtained by collecting the set of all `REFEX`s that are in the text, and adding the following defaults: for each `REFEX` that is a named reference in the genitive form, add the corresponding plain `REFEX`; conversely, for each `REFEX` that is a named reference not in the genitive form, add the corresponding genitive `REFEX`; for each `REFEX` that is a named reference add pronoun `REFEX`s of the appropriate number and gender, in the nominative, genitive and accusative forms, a relative pronoun `REFEX` in the nominative, and an empty `REFEX` (i.e. one with `REG08-TYPE="empty"`).[11]

`REF` elements that are embedded in `REFEX` elements contained in an `ALT-REFEX` list have an unspecified `MENTION` id (the '?' value). Furthermore, such `REF` elements have had their enclosed `REFEX` removed, i.e. they are 'empty'. For example:

```
<ALT-REFEX>
  ...
  <REFEX ENTITY="2" REG08-TYPE="common" CASE="plain">a friend of <REF ENTITY="1" MENTION="?"
     SEMCAT="person" SYNCAT="np" SYNFUNC="obj"></REF></REFEX>
  ...
</ALT-REFEX>
```

[11] Any resulting duplicates are removed.

```
<?xml version="1.0" encoding="utf-8"?>
<!DOCTYPE GREC-ITEM SYSTEM "genchal09-grec.dtd">
<GREC-ITEM>
<TEXT ID="15">
<TITLE>Alexander Fleming</TITLE>

<PARAGRAPH>
<REF ENTITY="0" MENTION="1" SEMCAT="person" SYNCAT="np" SYNFUNC="subj">
  <REFEX ENTITY="0" REG08-TYPE="name" CASE="plain">Sir Alexander Fleming</REFEX>
</REF>
(6 August 1881 - 11 March 1955) was a Scottish biologist and pharmacologist.
<REF ENTITY="0" MENTION="2" SEMCAT="person" SYNCAT="np" SYNFUNC="subj">
  <REFEX ENTITY="0" REG08-TYPE="name" CASE="plain">Fleming</REFEX>
</REF>
published many articles on bacteriology, immunology, and chemotherapy.
<REF ENTITY="0" MENTION="3" SEMCAT="person" SYNCAT="np" SYNFUNC="subj-det">
  <REFEX ENTITY="0" REG08-TYPE="pronoun" CASE="genitive">his</REFEX>
</REF>
best-known achievements are the discovery of the enzyme lysozyme in 1922 and the discovery
of the antibiotic substance penicillin from the fungus Penicillium notatum in 1928, for which
<REF ENTITY="0" MENTION="4" SEMCAT="person" SYNCAT="np" SYNFUNC="subj">
  <REFEX ENTITY="0" REG08-TYPE="pronoun" CASE="nominative">he</REFEX>
</REF>
shared the Nobel Prize in Physiology or Medicine in 1945 with
<REF ENTITY="1" MENTION="1" SEMCAT="person" SYNCAT="np" SYNFUNC="obj">
  <REFEX ENTITY="1" REG08-TYPE="name" CASE="plain">Florey</REFEX>
</REF>
and
<REF ENTITY="2" MENTION="1" SEMCAT="person" SYNCAT="np" SYNFUNC="obj">
  <REFEX ENTITY="2" REG08-TYPE="name" CASE="plain">Chain</REFEX>
</REF>
.</PARAGRAPH>
</TEXT>

<ALT-REFEX>
  <REFEX ENTITY="0" REG08-TYPE="empty" CASE="no_case">_</REFEX>
  <REFEX ENTITY="0" REG08-TYPE="name" CASE="genitive">Fleming's</REFEX>
  <REFEX ENTITY="0" REG08-TYPE="name" CASE="genitive">Sir Alexander Fleming's</REFEX>
  <REFEX ENTITY="0" REG08-TYPE="name" CASE="plain">Fleming</REFEX>
  <REFEX ENTITY="0" REG08-TYPE="name" CASE="plain">Sir Alexander Fleming</REFEX>
  <REFEX ENTITY="0" REG08-TYPE="pronoun" CASE="accusative">him</REFEX>
  <REFEX ENTITY="0" REG08-TYPE="pronoun" CASE="genitive">his</REFEX>
  <REFEX ENTITY="0" REG08-TYPE="pronoun" CASE="nominative">he</REFEX>
  <REFEX ENTITY="0" REG08-TYPE="pronoun" CASE="nominative">who</REFEX>
  <REFEX ENTITY="1" REG08-TYPE="empty" CASE="no_case">_</REFEX>
  <REFEX ENTITY="1" REG08-TYPE="name" CASE="genitive">Florey's</REFEX>
  <REFEX ENTITY="1" REG08-TYPE="name" CASE="plain">Florey</REFEX>
  <REFEX ENTITY="1" REG08-TYPE="pronoun" CASE="accusative">him</REFEX>
  <REFEX ENTITY="1" REG08-TYPE="pronoun" CASE="genitive">his</REFEX>
  <REFEX ENTITY="1" REG08-TYPE="pronoun" CASE="nominative">he</REFEX>
  <REFEX ENTITY="1" REG08-TYPE="pronoun" CASE="nominative">who</REFEX>
  <REFEX ENTITY="2" REG08-TYPE="empty" CASE="no_case">_</REFEX>
  <REFEX ENTITY="2" REG08-TYPE="name" CASE="genitive">Chain's</REFEX>
  <REFEX ENTITY="2" REG08-TYPE="name" CASE="plain">Chain</REFEX>
  <REFEX ENTITY="2" REG08-TYPE="pronoun" CASE="accusative">him</REFEX>
  <REFEX ENTITY="2" REG08-TYPE="pronoun" CASE="genitive">his</REFEX>
  <REFEX ENTITY="2" REG08-TYPE="pronoun" CASE="nominative">he</REFEX>
  <REFEX ENTITY="2" REG08-TYPE="pronoun" CASE="nominative">who</REFEX>
</ALT-REFEX>
</GREC-ITEM>
```

Fig. 3. Example text from GREC-People corpus.

Name:	*Values:*	*Explanation:*
ENTITY	integer	identifier for the discourse entity that is being referred to, unique within the text
MENTION	integer, ?	identifier for references to a given entity, unique for the entity
SEMCAT	person	the semantic category of the referent
SYNCAT	np, nom	the syntactic category required of referential expressions for the referent in the given context
SYNFUNC	subj	subject of a clause other than below
	subj_ger-part	subject of a gerund-participial ([17], p. 1191–1193)
	subj_rel-clause	subject within a relative clause
	obj	(in)direct object or object of PP other than below
	obj_rel-clause	object within a relative clause
	obj_refl	object of verb referring to same entity as subject of same verb
	app-supp	appositive supplement (e.g.: *George Sarton, the father of the history of science*)
	attr_compl	attributive complement (e.g.: *the Fessenden patents*)
	subj-det	genitive subject-determiner (e.g.: *his parents*)
	subj-det_rel-clause	genitive subject-determiner within a relative clause

Fig. 4. REF attribute names and values in GREC-People.

5 GREC-MSR and GREC-NEG Task Definitions

5.1 GREC-MSR

The training/development data in GRE-MSR is exactly as shown in Figure 2. The test data is the same, except that of course REF elements contain only an ALT-REFEX list, not the actual REFEX. The task for participating systems was to select one of the REFEXs in the ALT-REFEX list, for each REF in each TEXT in the test sets. The selected REFEX then had to be inserted into the REF in test set outputs submitted for the GREC-MSR Task.

In the first run of this task (part of REG'08[12]), the main task aim was to get the REG08-TYPE of selected referring expressions (REs) right, and REG08-Type Accuracy (for definition see Section 6) was the main evaluation metric. In the GREC-MSR'09 run of this task (part of GenChal'09[13]), the main task aim was to get the actual RE (the word string) right, and the main evaluation criterion was therefore Word String Accuracy.

We created four test sets for the GREC-MSR Task:

[12] The Referring Expression Generation Challenge 2008, see http://www.itri.brighton.ac.uk/research/reg08.

[13] Generation Challenges 2009, see http://www.itri.brighton.ac.uk/research/genchal09.

1. GREC-MSR Test Set C-1: a randomly selected 10% subset (183 texts) of the GREC corpus (with the same proportions of texts in the 5 subdomains as in the training/testing data).
2. GREC-MSR Test Set C-2: the same subset of texts as in C-1; however, for C-2 we did not use the REs in the corpus, but replaced them with human-selected alternatives. These were obtained in an online experiment (with an interface designed as shown in Figure 1) where participants selected REs in a setting that duplicated the conditions in which the participating systems in the GREC-MSR Task make selections.[14] We obtained three versions of each text, where in each version all REs were selected by the same person. The motivation for this version of Test Set C was that having several human-produced chains of REs against which to compare the outputs of participating ('peer') systems is more reliable than having one only; and that Wikipedia texts are edited by multiple authors which sometimes adversely affects MSR chains; we wanted to have additional reference texts where all references are selected by a single author.
3. GREC-MSR Test Set L: 74 Wikipedia introductory texts from the subdomain of lakes (there were no lake texts in the training/development set).
4. GREC-MSR Test Set P: 31 short encyclopaedic texts in the same 5 subdomains as in the GREC-2.0 corpus, in approximately the same proportions as in the training/testing data, but of different origin. We transcribed these texts from printed encyclopaedias published in the 1980s which are not available in electronic form. The texts in this set are much shorter and more homogeneous than the Wikipedia texts, and the sequences of MSRs follow very similar patterns. It seems likely that it is these properties that have resulted in better scores overall for Test Set P than for the other test sets in both the 2008 and 2009 runs of the GREC-MSR task.

Each test set was designed to test peer systems for generalisation to different kinds of unseen data. Test Set C tests for generalisation to unseen material from the same corpus and the same subdomains as the training set; Test Set L tests for generalisation to unseen material from the same corpus but different subdomain; and Test Set P for generalisation to a different corpus but the same subdomains.

5.2 GREC-NEG

The training/development data in GREC-NEG is exactly as shown in Figure 3. The test data is identical to the training/development data, except that `REF` elements do not contain a `REFEX` element, i.e. they are 'empty'.

The task is to select one `REFEX` from the `ALT-REFEX` list for each `REF` in each `TEXT` in the test sets. If the selected `REFEX` contains an embedded `REF` then participating systems also need to select a `REFEX` for this embedded `REF` and to set the value of the `MENTION` attribute (which has the '?' value in `REF`s that are embedded in

[14] The experiment can be tried out here: http://www.itri.brighton.ac.uk/home/Anja.Belz/TESTDRIVE/

Evaluation criterion	*Type of evaluation*	*Evaluation technique*
Humanlikeness	intrinsic/automatic	REG08-Type Accuracy, String Accuracy, String-edit distance, BLEU-3, NIST
Referential clarity	extrinsic/automatic	Automatic coreference resolution experiment
	intrinsic/human	Native speakers' judgement of clarity
Fluency	intrinsic/human	Native speakers' judgement of fluency
Ease of comprehension	extrinsic/human	Reading speed and comprehension accuracy measured in a reading experiment

Table 1. Overview of evaluation methods used in GREC Shared Task Evaluations.

`REFEX`s in `ALT-REFEX` lists). The same applies to all further embedded `REFEX`s, at any depth of embedding.

In the first run[15] of this task (part of GenChal'09, see Footnote 9), the main task aim was to get the `REG08-TYPE` of selected referring expressions (REs) right, and REG08-Type Accuracy (for definition see Section 6) was the main evaluation metric.

We created two versions of the test data for the GREC-NEG Task:

1. GREC-NEG Test Set 1a: randomly selected 10% subset (100 texts) of the GREC-People corpus (with the same proportion of texts in the 3 subdomains as in the training/development data).
2. GREC-NEG Test Set 1b: the same subset of texts as in (1a); for this set we did not use the REs in the corpus, but replaced each of them with human-selected alternatives obtained in an online experiment as for GREC-MSR.

6 GREC Evaluation Procedures

As in the TUNA evaluations (see Gatt & Belz elsewhere in this volume [15]), we developed a portfolio of intrinsic and extrinsic, human-assessed and automatically computed evaluation methods to assess the quality of the REs generated by GREC systems. Table 1 is an overview of the techniques we have used in the GREC evaluations. Each technique is explained in one of the subsections below.

In all GREC shared tasks, the data is divided into training, development and test data. In each case, we, the organisers, performed evaluations on the test data, using a range of different evaluation methods. Participants computed evaluation scores on the development set, using the `geval-2.0.pl` code provided by us which (in its most recent version) computes Word String Accuracy, REG'08-Type Recall and Precision, string-edit distance and BLEU.

[15] The second run was held in 2010 as part of GenChal'10, see http://www.itri.brighton.ac.uk/research/genchal10.

Some of the test sets have a single version of each text (the original corpus text, as in GREC-MSR test set C-1 and GREC-NEG test set 1a, see Sections 5.1 and 5.2), and the scoring metrics below that are based on counting matches (Word String Accuracy counts matching word strings, REG08-Type Accuracy, Recall and Precision count matching REG08-Type attribute values) simply count the number of matches a system achieves against that single text.

For each task we also created one test set which has three versions of each text with human-selected REs in them (C-2 in GREC-MSR and 1b in GREC-NEG). For these sets, the match-based metrics first calculate the number of matches for each of the three versions and then use (just) the highest number of matches in further calculations.

6.1 Automatic intrinsic evaluations of Humanlikeness

One set of humanlikeness measures we computed were REG08-Type Accuracy (GREC-MSR), and REG08-Type Recall and Precision (GREC-NEG). REG08-Type Precision is defined as the proportion of `REFEX`s selected by a participating system that mach the corresponding `REFEX`s in the evaluation corpus; REG08-Type Recall is defined as the proportion of `REFEX`s in the evaluation corpus for which a participating system has produced a match. For GREC-MSR Recall equals Precision, which is why we call it Accuracy there.

The reason why we use REG08-Type Recall and Precision for GREC-NEG rather than REG08-Type Accuracy as in GREC-MSR is that in GREC-NEG there may be a different number of `REFEX`s in system outputs and the reference texts in the test set (because there are embedded references in GREC-People, and systems may select `REFEX`s with or without embedded references for any given `REF`). In GREC-MSR, the number of `REFEX`s in a system output and the corresponding reference texts in the test set is the same, hence we compute just one score, REG08-Type Accuracy.

For both tasks, we computed String Accuracy, defined as the proportion of word strings selected by a participating system that match those in the reference texts. For GREC-NEG this was computed on the complete text within the outermost `REFEX`, including the text in embedded `REFEX` nodes.

We also computed BLEU-3, NIST, string-edit distance and length-normalised string-edit distance, all on word strings defined as for String Accuracy. As regards the tests sets with multiple REs, BLEU and NIST are designed for multiple output versions (so they could be applied as they are), whereas for the string-edit metrics we computed the mean of means over the three text-level scores (computed against the three versions of a text).

6.2 Automatic extrinsic evaluation of Clarity

In all three GREC shared-task evaluations, we used Coreference Resolver Accuracy (CRA), an automatic extrinsic evaluation method based on coreference

resolution performance. The basic idea is that it seems likely that badly chosen reference chains affect the ability to resolve REs in automatic coreference resolution tools.

To counteract the possibility of results being a function of a specific coreference resolution algorithm or evaluation method, we used several resolution tools and several evaluation methods and averaged results. There does not appear to be a single standard evaluation metric in the coreference resolution community. We opted to use the following three: MUC-6 [38], CEAF [23], and B-CUBED [1], which seem to be the most widely accepted metrics. All three metrics compute Recall, Precision and F-Scores on aligned gold-standard and resolver-tool coreference chains. They differ in how the alignment is obtained and which components of coreference chains are counted for calculating scores.

In GREC-MSR'08, we used three different resolvers—those included in LingPipe,[16] JavaRap [30] and OpenNLP [24]. However, for GREC'09 we overhauled the CRA tool; the current version no longer uses JavaRAP, and uses the most recent versions of the other resolvers; the GREC-MSR'08 and GREC-MSR'09 results for this method are not entirely comparable for this reason.

For each system, the CRA tool runs the coreference resolvers on each system output, then CRA computes the MUC-6, CEAF and B-CUBED F-Scores for each coreference resolver output, then their mean, and finally the mean over all system outputs.

6.3 Human-assessed intrinsic evaluation of Clarity and Fluency

6.3.1 GREC-MSR'09

The intrinsic human evaluation in GREC-MSR'09 involved 24 randomly selected items from Test Set C and outputs for these produced by peer and baseline systems (described in Section 7.1) as well as those found in the original corpus texts (8 'systems' in total). We used a Repeated Latin Squares design which ensures that each participant sees the same number of outputs from each system and for each test set item. There were three 8×8 squares, and a total of 576 individual judgements in this evaluation (72 per system: 3 criteria $\times 3$ articles $\times 8$ evaluators).

We recruited 8 native speakers of English from among post-graduate students currently doing a linguistics-related degree at University College London (UCL) and Sussex University.

Following detailed instructions, participants did two practice examples, followed by the 24 texts to be evaluated, in random order. Subjects carried out the evaluation over the internet, at a time and place of their choosing. They were allowed to interrupt and resume the experiment (though discouraged from doing so). According to self-reported timings, participants took between 25 and 45 minutes to complete the evaluation (not counting breaks).

[16] `http://alias-i.com/lingpipe/`

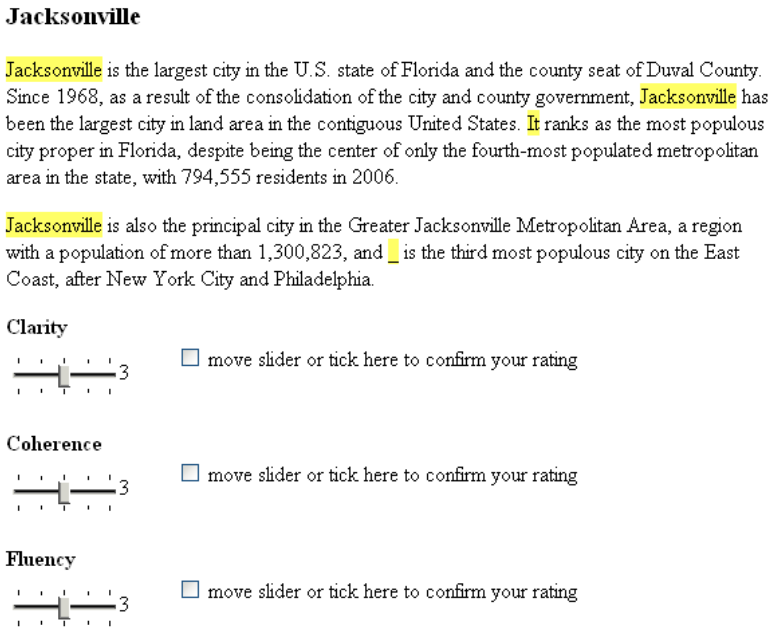

Fig. 5. Example of text presented in human intrinsic evaluation of GREC-MSR systems.

Figure 5 shows what participants saw during the evaluation of an individual text. All references to the MS were highlighted in yellow,[17] and the task is to evaluate the quality of the REs in terms of three criteria which were explained in the introduction as follows (the wording of the explanations of Criteria 1 and 3 were taken from the DUC evaluations):

1. **Referential Clarity**: It should be easy to identify who or what the referring expressions in the text are referring to. If a person or other entity is mentioned, it should be clear what their role in the story is. So, a reference would be unclear if an entity is referenced, but their identity or relation to the story remains unclear.
2. **Fluency**: A referring expression should 'read well', i.e. it should be written in good, clear English, and the use of titles and names etc. should seem natural. Note that the Fluency criterion is independent of the Referential Clarity criterion: a reference can be perfectly clear, yet not be fluent.
3. **Structure and Coherence**: The text should be well structured and well organised. The text should not just be a heap of related information, but should build from sentence to sentence to a coherent body of information about a topic. This criterion too is independent of the others.

[17] Showing up as pale shaded boxes around words in black and white versions of this document.

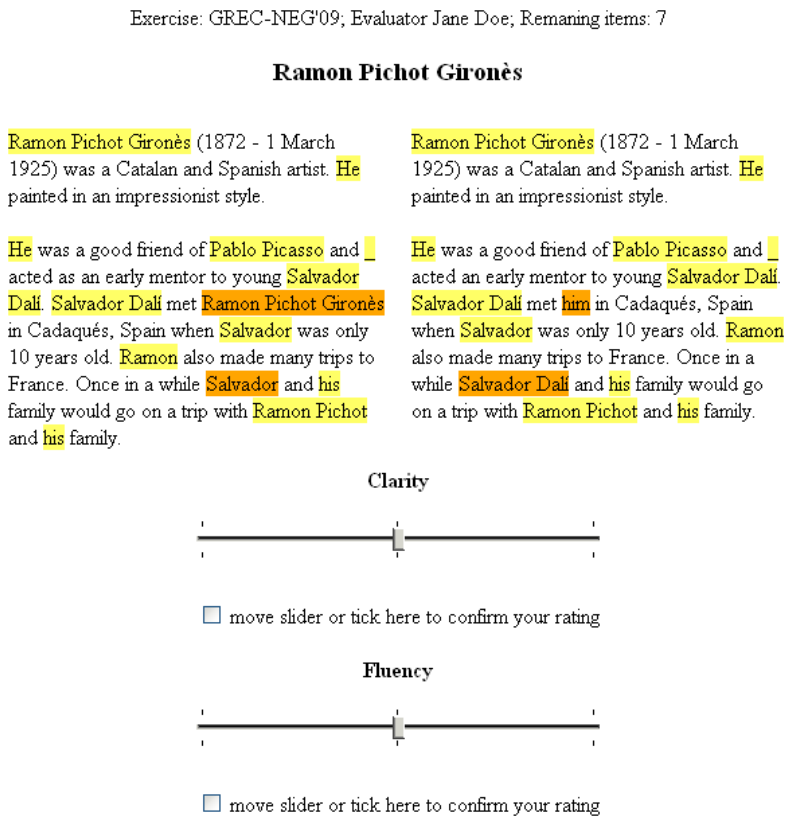

Fig. 6. Example of text pair presented in human intrinsic evaluation of GREC-NEG systems.

Subjects selected evaluation scores by moving sliders (see Figure 5) along scales ranging from 1 to 5. Slider pointers started out in the middle of the scale (3). These were continuous scales and we recorded scores with one decimal place (e.g. 3.2). The meaning of the numbers was explained in terms of integer scores (1=very poor, 2=poor, 3=neither poor nor good, 4=good, 5=very good).

6.3.2 GREC-NEG'09

The motivating application context for the GREC-NEG Task is, as mentioned above (Section 2.1), improving referential clarity and coherence in multiply edited texts. We therefore designed the human-assessed intrinsic evaluation as a preference-judgement test where participants expressed their preference, in terms of two criteria, for either the original Wikipedia text or the version of it

with system-generated referring expressions in it. The intrinsic human evaluation involved outputs for 30 randomly selected items from the test set from 5 of the 6 participating systems,[18] four baselines and the original corpus texts (10 'systems' in total, all described in Section 8.1). Again, we used a Repeated Latin Squares design. This time there were three 10×10 squares, and a total of 600 individual judgements in this evaluation (60 per system: 2 criteria $\times 3$ articles $\times 10$ evaluators). We recruited 10 native speakers of English from among students currently completing a linguistics-related degree at Kings College London and University College London.

As in the GREC-MSR and TUNA evaluation experiments, participants were given detailed instructions, two practice examples, and then the texts to be evaluated, in random order. Subjects did the evaluation over the internet, at a time and place of their choosing, and were allowed (though discouraged) to interrupt and resume.

Figure 6 shows what participants saw during the evaluation of an individual text pair. The place (left/right) of the original Wikipedia article was randomly determined for each individual evaluation. People references are highlighted in yellow/orange, those that are identical in both texts are yellow, those that are different are orange.[19] The evaluator's task is to express their preference, as well as the strength of their preference, in terms of each quality criterion by moving the slider pointers. Moving the slider to the left means expressing a preference for the text on the left, moving it to the right means preferring the text on the right; the further to the left/right the slider is moved, the stronger the preference. The two criteria of Fluency and Referential Clarity were explained to participants in the introduction with exactly the same wording as described above for GREC-MSR (Section 6.3.1).

In this experiment, unlike in the GREC-MSR experiment, it was not evident to the evaluators that sliders were associated with numerical values. Slider pointers started out in the middle of the scale (no preference). The values associated with the points on the slider ranged from -10.0 to +10.0.

6.4 Human-based extrinsic evaluation of Ease of Comprehension

For GREC-MSR'08, we designed a reading/comprehension experiment in which the task for participants was to read texts one sentence at a time and then to answer three brief multiple-choice comprehension questions after reading each text. The basic idea was that it seemed likely that badly chosen MSR reference chains would adversely affect ease of comprehension, and that this might in turn affect reading speed.

We used a randomly selected subset of 21 texts from GREC-MSR Test Set C, and recruited 21 participants from among the staff, faculty and students of

[18] We left out UDel-NEG-1 given our limited resources and the fact that this is a kind of baseline system.

[19] When viewed in black and white, the orange highlights appear slightly darker than the yellow ones.

Brighton and Sussex universities. We used a Repeated Latin Squares design in which each combination of text and system was allocated three trials. During the experiment we recorded *SRTime* (sentence reading time), the time participants took to read sentences (from the point when the sentence appeared on the screen to the point at which the participant requested the next sentence).

We also recorded the speed and accuracy with which participants answered the questions at the end (*Q-Time* and *Q-Acc*). The role of the comprehension questions was to encourage participants to read the texts properly, rather than skimming through them, and we did not necessarily expect any significant results from the associated measures.

The questions were designed to be of varying degrees of difficulty and predictability. There were three questions (each with five possible answers) associated with each text, and questions followed the same pattern across the texts: the first question was always about the subdomain of a text (*The text I just read was about a [city/country/river/person/mountain]*); the second about the geographical location of the main subject (e.g. *The city I just read about is located in [Peshawar/Uttar Pradesh/...]*; *The person I just read about was born in [England/Scotland/...]*); the third question was designed not to be predictable (e.g. *How many hydroelectric power stations are there on this river? [three/five/four/...]*; *This mountain is the location of a neolithic [jadeite quarry/ jasper quarry/...]*).

The order of the possible answers was randomised for each question and each participant. The order of texts (with associated questions) was randomised for each participant. We used the DMDX package for presentation of sentences and measuring reading times and question answering accuracy [14]. Subjects did the experiment in a quiet room, under supervision.

7 GREC-MSR'08/09—Participating Systems and Results

7.1 Systems

In this section, we give very brief descriptions of the systems that participated in the two GREC-MSR competitions. Full details can be found in the reports from participating teams in the INLG'08 proceedings (for GREC-MSR'08) and the ENLG'09 proceedings (for GREC-MSR'09).

Base-rand, Base-freq, Base-1st, Base-name: We created four baseline systems. *Base-rand* selects one of the `REFEX`s at random. *Base-freq* selects the `REFEX` that is the overall most frequent given the `SYNCAT` and `SEMCAT` of the reference. *Base-1st* always selects the `REFEX` which appears first in the list of `REFEX`s; and *Base-name* selects the shortest `REFEX` with attributes `REG08-TYPE=name`, `HEAD=nominal` and `EMPHATIC=no`.[20]

[20] Attributes are tried in this order. If for one attribute, the right value is not found, the process ignores that attribute and moves on the next one.

CNTS-Type-g, CNTS-Prop-s (GREC-MSR'08) : The CNTS systems are trained using memory-based learning with automatic parameter optimisation. They use a set of 14 features obtained by various kinds of syntactic preprocessing and named-entity recognition as well as from the corpus annotations: `SEMCAT`, `SYNCAT`, position of RE in text, neighbouring words and POS-tags, distance to previous mention, `SYNCAT`s of the three preceding `REFEX`s, a binary feature indicating whether the most recent named entity was the main subject (entity), and the main verb of the sentence. For *CNTS-Type-g*, a single classifier was trained to predict just the `REG08-TYPE` property of `REFEX`s. For *CNTS-Prop-s*, four classifiers were trained, one for each subdomain, to predict all four properties of `REFEX`s (rather than just `REG08-TYPE`).

OSU-b-all, OSU-b-nonRE, OSU-n-nonRE (GREC-MSR'08): The *OSU* systems are maximum-entropy classifiers trained on a range of features obtained from the corpus annotations and by preprocessing the text: `SEMCAT`, `SYNCAT`, position of RE in text, presence of contrasting discourse entity, distance between current and preceding reference to the entity, string similarity measures between `REFEX`s and the title of text. *OSU-b-all* and *OSU-b-nonRE* are binary classifiers which give the likelihood of selecting a given `REFEX` vs. not selecting it, whereas *OSU-n-nonRE* is a 4-class classifier giving the likelihoods of selecting each of the four `REG08-TYPE`s. *OSU-b-all* also uses the `REFEX` attributes as features.

IS-G (GREC-MSR'08): The *IS-G* system is a multi-layer perceptron which uses four features obtained by preprocessing texts and from the corpus annotations: `SYNCAT`, distance between current and preceding reference to the entity, position of RE in text, `REG08-TYPE` of preceding reference to the entity, feature indicating whether the preceding mention is in the same sentence.

UDel (GREC-MSR'09): The *UDel* system is informed by psycholinguistic research and consists of a preprocessing component performing sentence segmentation and identification of non-referring occurrences of entity names, an RE type selection component (two C5.0 decision trees, one optimised for people and mountains, the other for the other subdomains), and a word string selection component. The RE type selection decision trees use the following features: binary features indicating whether the entity is the subject of the current, preceding and preceding but one sentences, whether the last MSR was in subject position, and whether there are intervening references to other entities between the current and the previous MSR. Other features encode distance to preceding non-referring occurrences of an entity name; sentence and reference IDs; and whether the reference occurred before and after certain words and punctuation marks. Given a selected RE type, the word-string selection component selects (from among those `REFEX`s that have a matching type) the longest non-emphatic name for the first named reference in an article, and the shortest for subsequent named references; for other types, the first matching word-string is used, backing off to pronoun and name.

System	REG08-Type Accuracy						WSAcc	BLEU-3	NIST	SE	SRTime	Q1	CRA
Corpus	78.58	A					71.18	0.779	7.508	0.723	6548	1.00	43.32
CNTS-Type-g	72.61	A	B				65.61	0.738	6.129	0.884	6436	1.00	48.64
CNTS-Prop-s	71.34		B				65.29	0.676	5.934	0.907	6306	0.98	46.35
IS-G	70.78		B				58.20	0.511	5.610	1.162	6341	0.95	48.05
OSU-n-nonRE	69.82		B				63.85	0.672	5.775	0.967	6423	0.90	51.39
OSU-b-nonRE	58.76			C			51.11	0.496	5.536	1.283	6455	0.97	51.27
OSU-b-all	57.48			C			50.72	0.505	5.606	1.299	6452	0.95	50.87
Base-name	50.00				D		39.41	0.464	5.937	1.518	–	–	52.84
Base-1st	49.28				D		39.09	0.393	5.16	1.645	–	–	53.50
Base-freq	48.17				D		41.32	0.268	3.016	1.543	–	–	41.41
Base-rand	41.24					E	17.99	0.218	2.932	2.322	–	–	35.13

Table 2. GREC-MSR'08: REG08-Type Accuracy scores with homogeneous subsets (Tukey HSD, alpha = .05), string-similarity scores, sentence reading time (SRT), Question 1 Accuracy (Q1), and coreference resolver accuracy (CRA), automatic metrics as computed against Test Set C-2.

ICSI-CRF (GREC-MSR'09): The *ICSI-CRF* system construes the GREC-MSR task as a sequence labelling task and determines the most likely current label given preceding labels, using a Conditional Random Field model trained on the following features for the current, preceding and preceding but one MSR: preceding and following word unigram and bigram; suffix of preceding and following word; preceding and following punctuation; reference ID; and a binary feature encoding whether the current sentence is the beginning of a paragraph. If more than one label remains, the last in the list of possible REs in the GREC-MSR data is selected.

JUNLG (GREC-MSR'09): The *JUNLG* system is based on co-occurrence statistics between `REF` feature sets and `REFEX` feature sets as found in the GREC-MSR data. `REF` feature sets were augmented by a paragraph counter and a within-paragraph `REF` counter. For each given set of `REF` features, the system selects the most frequent `REFEX` feature set (as determined from co-occurrence counts in the training data). If the current set of possible `REFEX`s does not include a `REFEX` with the selected feature set, then the second most likely feature set is selected. Several hand-coded default rules override the frequency-based selections, e.g. if the preceding word is a conjunction, and the current `SYNCAT` is np-subj, then the REG08-Type is empty.

7.2 Results

Table 2 shows mean scores for all evaluation methods in GREC-MSR'08; in the case of the automatically computed metrics, scores were computed against Test Set C-2, which has 3 versions of each text in the test set, each with people REs selected by a (different) human. Statistically significant differences are indicated (in the form of homogeneous subsets) only for REG08-Type Accuracy which was the nominated main evaluation method of GREC-MSR'08.[21]

[21] By 'nominated main evaluation method' we mean that this was the method that participants were told was going to be the main method by which systems would

	REG08-T. Acc	WSAcc	BLEU-3	NIST	SE	SRTime	Q1	CRA
REG08Type Accuracy	1	.964**	.934**	.795**	-.937**	.045	.334	.201
WSAcc	.964**	1	.934**	.802**	-.994**	.120	.332	.341
BLEU-3	.934**	.934**	1	.896**	-.932**	.289	.396	.353
NIST	.795**	.802**	.896**	1	-.822**	.616	.553	.545
SE	-.937**	-.994**	-.932**	-.822**	1	-.199	-.398	-.390
SRTime	.045	.120	.289	.616	-.199	1	.241	-.140
Q1 Accuracy	.334	.332	.396	.553	-.398	.241	1	-.656
CRA	.201	.341	.353	.545	-.390	-.140	-.656	1

Table 3. GREC-MSR'08: Pearson's correlation coefficients for all evaluation methods in Table 2. **. Correlation is significant at the 0.01 level (2-tailed).

Table 3 shows the corresponding Pearson's correlation coefficients for all evaluation methods in Table 2. The picture is very clear: all corpus-similarity metrics (whether based on string similarity or RE type similarity) are strongly and highly significantly correlated with each other. However they are not correlated significantly with any of the extrinsic methods, and there are also no significant correlations between any of the extrinsic methods.[22]

Table 4 shows analogous results for GREC-MSR'09. This time, the table shows statistical significance for Word String Accuracy, as this was the nominated main evaluation method for GREC-MSR'09. Table 5 shows the corresponding Pearson's correlation coefficients. This time the picture is not quite so simple. Once again, the automatically computed corpus-similarity metrics correlate strongly and highly significantly with each other. Out of the human-assessed intrinsic metrics, Fluency and Coherence correlate strongly with the automatically computed corpus-similarity metrics. However, Clarity only correlates with NIST and (to a lesser extent) with SE. While Fluency and Coherence correlate well with each other, only Coherence is also correlated with Clarity. The correlation between Fluency and Coherence makes sense intuitively, since both could be seen as dimensions of how well a text reads. The weaker correlations with Clarity indicate that the human evaluators were able to consider it to some degree independently from the other criteria, and there must have been some systems that produced texts that were clear but not fluent (or even vice versa). The slightly stronger correlation between Coherence and Clarity also makes sense, since, for example, using pronouns in the right place contributes to both (but not necessarily to Fluency).

be evaluated. The reason why this is important is that some participants may have optimised their systems for the nominated main evaluation method.

[22] For comparison with a similar lack of correlations between intrinsic and extrinsic methods in the TUNA tasks, see the discussion section below (Section 9).

System	REG08-T. Acc.	Word String Accuracy							BLEU-3	NIST	SE	Cla	Flu	Coh	CRA
Corpus	79.30	71.58	A						0.77	5.60	1.04	4.56	4.43	4.40	42.52
UDel	77.71	70.22	A	B					0.74	5.32	1.11	4.35	4.27	4.27	46.19
JUNLG	75.40	64.57		B	C				0.53	4.69	1.34	4.50	4.26	4.33	44.19
ICSI-CRF	75.16	63.69			C				0.54	4.68	1.32	4.45	4.15	4.02	44.47
Base-freq	62.50	57.01				D			0.54	4.30	1.93	4.10	3.33	3.96	63.14
Base-name	51.04	40.21					E		0.46	4.76	1.80	4.62	2.84	3.85	65.19
Base-1st	50.32	39.65					E		0.39	4.42	1.93	4.27	2.76	3.7	63.77
Base-rand	48.09	26.99						F	0.26	3.02	2.30	3.18	2.15	3.46	42.99

Table 4. GREC-MSR'09: REG08-Type Accuracy scores, Word String Accuracy with homogeneous subsets (Tukey HSD, alpha = .05), other string-similarity scores, Clarity, Fluency, Coherence scores, and coreference resolver accuracy (CRA), automatic metrics as computed against Test Set C-2.

	REG08-T. Acc.	WSA	BLEU-3	NIST	SE	Cla	Flu	Coh	CRA
REG08-Type Acc.	1	.971**	.862**	.726*	-.931**	.531	.984**	.922**	-.609
Word String Acc.	.971**	1	.923**	.818*	-.925**	.645	.983**	.950**	-.407
BLEU3	.862**	.923**	1	.909**	-.905**	.649	.881**	.909**	-.293
NIST	.726*	.818*	.909**	1	-.891**	.880**	.812*	.860**	-.073
SE	-.931**	-.925**	-.905**	-.891**	1	-.717*	-.959**	-.930**	.488
Clarity	.531	.645	.649	.880**	-.717*	1	.670	.714*	.181
Fluency	.984**	.983**	.881**	.812*	-.959**	.670	1	.956**	-.493
Coherence	.922**	.950**	.909**	.860**	-.930**	.714*	.956**	1	-.386
Coref. Resolver Acc.	-.609	-.407	-.293	-.073	.488	.181	-.493	-.386	1

Table 5. GREC-MSR'09: Pearson's correlation coefficients for all evaluation methods in Table 4. **. Correlation is significant at the 0.01 level (2-tailed).

8 GREC-NEG'09—Participating Systems and Results

8.1 Systems

In this section, we give very brief descriptions of the systems that participated in the GREC-NEG competition. Individual reports describing participating systems can be found in the proceedings of the UCNLG+SUM workshop.

Base-rand, Base-freq, Base-1st, Base-name: We created four baseline systems each with a different way of selecting a `REFEX` from those `REFEX`s in the `ALT-REFEX` list that have matching entity IDs. *Base-rand* selects a `REFEX` at random. *Base-1st* selects the first `REFEX`. *Base-freq* selects the first `REFEX` with a `REG08-TYPE` that is the overall most frequent (as determined from the training/development data) given the `SYNCAT`, `SYNFUNC` and `SEMCAT` of the reference. *Base-name* selects the shortest `REFEX` with attribute `REG08-TYPE=name`.

UDel-NEG-1, UDel-NEG-2, UDel-NEG-3: The *UDel-NEG-1* system is identical to the *UDel* system that was submitted to the GREC-MSR'09 Task (for a description of that system see Section 7.1 above), except that it was adapted to the different data format of GREC-NEG. *UDel-NEG-2* is identical to *UDel-NEG-1* except that it was retrained on GREC-NEG data and the feature set was extended by entity and mention IDs. *UDel-NEG-3* additionally utilised improved identification of other entities.

System	REG08-Type														WSAcc	BLEU-3	NIST	nSE	Cla	Flu	CRA
	Precision							Recall													
Corpus	82.67	A						84.01	A						81.90	0.95	7.15	0.25	0	0	59.56
ICSI-CRF	79.33	A	B					78.38		B					74.69	0.86	6.36	0.31	-1.45	-0.35	61.28
WLV-BIAS	77.78		B					77.78		B					69.14	0.88	6.18	0.36	-2.44	-2.26	62.64
WLV-STAND	67.51			C				67.51			C				59.84	0.83	5.82	0.45	-4.48	-5.82	51.69
Base-freq	65.38			C				64.37			C				3.24	0.39	2.1	0.90	-8.26	-7.57	55.85
UDel-NEG-2	57.39				D			56.06				D			18.96	0.53	2.42	0.83	-6.67	-7.13	55.9
UDel-NEG-3	57.25				D			55.92				D			18.89	0.53	2.49	0.82	-6.43	-6.26	56.13
Base-name	55.22				D			54.01				D			37.27	0.65	5.57	0.63	-2.58	-4.26	61.11
UDel-NEG-1	53.57				D			52.32				D	E		19.25	0.51	2.62	0.82	–	–	54.79
Base-rand	48.46					E		47.75					E		10.45	0.25	1.11	0.89	-8.18	-7.51	34.86
Base-1st	12.54						F	12.54						F	8.65	0.24	1.29	0.92	-9.36	-8.48	26.36

Table 6. GREC-NEG'09: REG08-Type Recall and Precision scores with homogeneous subsets (Tukey HSD, alpha = .05), and all word-string similarity based intrinsic automatic scores. All scores as computed against human topline version of Test Set (1b).

ICSI-CRF: see Section 7.1.

WLV-BIAS, WLV-STAND: The *WLV* systems start with sentence splitting and POS tagging. *WLV-STAND* then employs a J48 decision tree classifier to obtain a probability for each `REF/REFEX` pair that it is a good pair in the current context. The context is represented by the following set of features. Features of the `REFEX` word string include features encoding whether the string is the longest of the possible `REFEX`s, the number of words in the string, and all `REFEX` features supplied in the GREC-NEG data. Features of the `REF` include features encoding whether the current entity's chain is the first in the text, whether the current mention of the entity is the first, whether the mention is at the beginning of a sentence, and all `REF` features supplied in GREC-NEG data. Other features include one that encodes whether the current `REF` is preceded by one of a given set of strings including *", but", "and then"* and similar phrases, the distance in sentences to the last mention, the REG08-Types selected for the two preceding `REF`s, the POS tags of the preceding four words and the following three words, the correlation between `SYNFUNC` and `CASE` values, and the size of the chain.

WLV-BIAS is the same except that it is retrained on reweighted training instances. The reweighting scheme assigns a cost of 3 to false negatives and 1 to false positives.

8.2 Results

Table 6 shows mean scores for all evaluation methods used in GREC-NEG'09; in the case of the automatically computed metrics, scores were computed against Test Set 1b, which has 3 versions of each text in the test set, each with people REs selected by a (different) human. Statistically significant differences (in the form of homogeneous subsets) are shown only for REG08-Type Precision and Recall which were the nominated main evaluation methods of GREC-NEG'09.

	REG08-T. Precision	REG08-T. Recall	WSAcc	BLEU-3	NIST	norm. SE	Clarity	Fluency	CRA
REG08-Type Prec.	1	.999**	.733*	.825**	.749**	-.760**	.765**	.762*	.844**
REG08-Type Rec.	.999**	1	.751**	.836**	.763**	-.777**	.776**	.773**	.831**
Word String Acc.	.733*	.751**	1	.952**	.957**	-.997**	.933**	.925**	.556
BLEU-3	.825**	.836**	.952**	1	.964**	-.954**	.934**	.879**	.753**
NIST	.749**	.763**	.957**	.964**	1	-.968**	.964**	.899**	.678*
norm. SE	-.760**	-.777**	-.997**	-.954**	-.968**	1	-.940**	-.932**	-.582
Clarity	.765**	.776**	.933**	.934**	.964**	-.940**	1	.958**	.734*
Fluency	.762*	.773**	.925**	.879**	.899**	-.932**	.958**	1	.664*
Coref. resolver Acc.	.844**	.831**	.556	.753**	.678*	-.582	.734*	.664*	1

Table 7. Pearson's correlation coefficients for all evaluation methods in Table 6. **= Correlation is significant at the 0.01 level (2-tailed). *= Correlation is significant at the 0.05 level (2-tailed).

Table 7 shows the corresponding correlation results (again for Pearson's r) at the system level. All intrinsic methods, both automatically computed and human-assessed, correlate well. Unlike in the GREC-MSR correlation results, here, coreference resolver accuracy (CRA) correlates well with all intrinsic methods except Word String Accuracy and string-edit distance. The strongest correlation is with REG08-Type Precision and Recall.

9 Discussion

9.1 Evaluation Methods

An important research focus in both the TUNA and the GREC shared-task challenges has been the evaluation methods we have used. Our dual aims have been to develop new evaluation methods, and to assess the performance of both existing and new evaluation methods.

One very experimental evaluation metric we developed is Coreference Resolver Accuracy (CRA). The original implementation included one resolver, JavaRAP, that was with hindsight not suitable, because it performs anaphora resolution rather than coreference resolution (and not all references are anaphoric). The removal of JavaRAP may explain the differences in correlation results between GREC-MSR'08 and GREC-MSR'09 (Tables 3 and 5): in the former, correlations tended to go in a positive direction, whereas in the latter, they tended to be in the negative direction. However, the only statistically significant correlation results are from GREC-NEG'09, where correlations (Table 7) with the intrinsic metrics were all in the 'right' direction (i.e. better intrinsic scores also implied better CRA). A clear bias of CRA is that it favours systems that produce a lot of named references (as these are easy for a coreference resolver to identify). CRA remains a highly experimental evaluation metric, but as both coreference resolution methods and named entity generation methods improve, it may become a viable evaluation method.

Another experimental evaluation method we developed is extrinsic evaluation by reading/comprehension experiments. The intuition here was that poorly

chosen references would slow down reading speed and interfere with text comprehension. The one time we ran this experiment (GREC-MSR'08), the differences in reading speeds between the participating systems were very small (and no statistically significant differences were found). We did (unexpectedly) find a significant (albeit weak) impact on the comprehension question that asked readers what the article they had just read was about (possible answers were person, city, country, river or mountain).

Of course, we cannot conclude from the lack of statistically significant differences between reading times that there are no differences to be found, and a different experimental design may well reveal such significant differences. For example, a contributing factor to our lack of results may have been that measuring reading time at the sentence level results in measurements on the order of seconds, and there is a lot of variance in such long reading times. We are planning to run this experiment again in the next GREC-NEG evaluation. This time, we will aim to measure reading time on smaller units of text (e.g. individual referential expressions), possibly using eye-tracking; we will also include the baseline systems. For the time being, however, this also remains a very experimental evaluation method.

Much more successful in the immediately term was the preference judgement experiment in GREC-NEG'09. Rating-scale evaluations, where human evaluators assess system outputs by selecting a score on a discrete scale, are the most common form of human-assessed evaluation in NLP, but are problematic for several reasons. Rating scales are unintuitive to use; deciding whether a given text deserves a 5, a 4 or a 3 etc. can be difficult. Furthermore, evaluators may ascribe different meanings to scores and the distances between them. Individual evaluators have different tendencies in using rating scales, e.g. what is known as 'end-aversion' tendency where certain individuals tend to stay away from the extreme ends of scales; other examples are positive skew and acquiescence bias, where individuals make disproportionately many positive or agreeing judgements (see e.g. [8]). It is not surprising then that stable averages of quality judgements, let alone high levels of agreement, are hard to achieve, as has been observed for MT [36, 22], text summarisation [35], and language generation [2]. The result of a rating scale experiment is ordinal data (sets of scores selected from the discrete rating scale). The means-based ranks and statistical significance tests that are commonly presented with the results of RSEs are not generally considered appropriate for ordinal data in the statistics literature [33]. At a minimum, "a test on the means imposes the requirement that the measures must be additive, i.e. numerical" [33, p. 14]. Parametric statistics are more powerful than non-parametric alternatives, because they make a number of strong assumptions (including that the data is numerical). If the assumptions are violated then the risks is that the significance of results is overestimated.

In view of these problems, we wanted to try out a preference judgement experiment, where evaluators compare two outputs at a time and simply have to state which of the two they think is better in terms of a given quality criterion. The experiment we designed (described in Section 6.3) added a twist to this

in that we used sliders with which participants could express the *strength* of their preference (see Figure 6). In all cases, participants were comparing system outputs to the original Wikipedia texts. Viewed in terms of binary preference decisions, this experiment gave very clear-cut results: all but the top two systems were *dispreferred* almost all of the time; the second ranked system was dispreferred and *neither preferred nor dispreferred* about the same number of times (and almost never preferred); the top ranking system (ICSI-CRF) was *dispreferred* only about a third of the time, and in the case of of Fluency, it was preferred more often than it was dispreferred. This last result indicates that the ICSI-CRF system actually succeeded in improving over the Wikipedia texts.

There was also less variance and more statistically significant differences in the results than in comparable rating-scale experiments (including evaluation experiments that use slider bars to represent rating scales such as in GREC-MSR'09 and TUNA'09). The correlations between Fluency and Clarity on the one hand and the string-similarity metrics on the other were very high (mostly above 0.92). The results are easy to interpret—strength of preference and preference counts are intuitive concepts. On the whole, preference judgements using slider scales with results reported in the above manner are a very promising new evaluation method for NLG, and we are currently conducting follow-up experiments to investigate it further.

In assessing the different evaluation methods we have used, we look at variance, how many statistically significant differences were found, and how intuitive the results are. After GREC-MSR'08, we decided not to use the ROUGE metrics any more (part of the problem was the infeasibly high rank both ROUGE-SU4 and ROUGE-2 assigned to the random baseline). As mentioned above, we continue to consider CRA and reading-comprehension experiments unreliable methods, albeit ones with potential from further development.

However, our main tool is looking at correlation coefficients between sets of system-level evaluation scores. In GREC-MSR'08, there were strong correlations between all pairs of intrinsic evaluation metrics, and none between any intrinsic and extrinsic measures. But the latter result is hard to interpret, because of the highly experimental nature of the extrinsic measures. Furthermore, very few statistically significant results were obtained for the extrinsic measures, so for the time being we do not know how meaningful they are.

In GREC-MSR'09, there were strong correlations between all pairs of intrinsic evaluation metrics (human-assessed and automatically computed) except for Clarity which only correlated significantly with NIST, SE and Coherence. The correlations with CRA were not significant.

Finally, in GREC-NEG'09, virtually all measures correlated strongly with each other; even CRA correlated well with most measures (except for Word String Accuracy and string-edit distance).

While the lack of correlation between extrinsic and intrinsic measures in GREC-MSR'08 and GREC-MSR'09 echoes similar results from the TUNA evaluations, we would caution against drawing any conclusions from this, because of the experimental nature of the extrinsic metrics. The one time we did get plenty

of significant differences for CRA (in GREC-NEG'09), we also had good correlation with the intrinsic measures.

Repeated comparative evaluation experiments can reveal patterns in how measures relate to each other. They can also show individual measures to yield inconsistent and/or unintuitive results over time. In the case of the remaining measures—those that do yield consistent and plausible results over time—two measures that consistently correlate highly in different experiments and for different tasks, could conceivably substitute for each other. But what we have avoided is to interpret any one of our measures as a basis for validating other measures. In MT and summarisation evaluation, good correlation with human judgements is often taken as validating an automatic metric (and conversely, lack of correlation as invalidating it). But intrinsic human judgements are simply not consistent and reliable enough to provide an objective meta-evaluation tool.[23] Moreover, all they provide is an insight into what humans (think they) like, not what is best or most useful for them (the two can be two very different matters, as discussed in [4]).

Ultimately, what those evaluation measures that are in themselves consistent, but do not consistently correlate strongly with each other, can provide us with are assessments of two different aspects of system quality. If there is anything we have learned from the numerous evaluation experiments we have carried out for TUNA and GREC, it is that a range of intrinsic and extrinsic methods, human-assessed and automatically computed, need to be applied in order to obtain a comprehensive view of system quality.

9.2 Development of the GREC Tasks

The GREC tasks are designed to build on each other, and to become increasingly complex and application-oriented. Our aims in this have been (i) to create data and define tasks that will enable REG researchers to investigate the REG problem as grounded in discourse context and informed by naturally occurring data; (ii) to build bridges to the summarisation and named-entity recognition communities by designing tasks that should be of specific interest to those researchers; and (iii) to move towards being able to use REG techniques in real application contexts.

Until recently, connections and joint forums between NLG and other fields in which language is automatically generated (notably MT and summarisation) were steadily decreasing as these fields moved away from language generation techniques and towards data-driven, direct text-to-text mappings. This made it difficult for NLG researchers to engage with these fields effectively and contribute to developing solutions for application tasks.

Grammar-based language generation techniques are making a comeback in MT and summarisation, as stastistical MT researchers have moved towards incorporating syntactic knowledge (e.g. under the heading of syntax-based statistical MT as in work by Knight and colleagues [6, 13]), and researchers in extractive

[23] Agreement among human judges is hard to achieve as has been discussed for MT, summarisation and NLG [36, 22, 35, 2].

summarisation have started to develop post-processing text regeneration techniques to improve the coherence and clarity of summaries [27, 34, 25].

This means that there is now more common ground than there has been in a long time between NLG and neighbouring fields, providing a good opportunity to bring researchers from the different fields together in working on intersecting tasks such as the GREC tasks.

We have already succeeded in attracting researchers from the Named Entity Recognition (NER) and machine learning communities as participants in GREC-NEG'09. In GREC'10, one of the subtasks is NER and another is an end to end RE regeneration task for extractive summarisation. The latter task will also be the first task to require stand-alone application systems to be developed.

10 Concluding Remarks

Participation in the GREC tasks has so far not reached the high levels of participation seen in the first two years of the TUNA tasks. This may be because the GREC tasks are entirely new research tasks, whereas TUNA involved tasks that many researchers were already working on (the first GIVE challenge also attracted just three teams from outside the team of organisers; see Koller et al. elsewhere in this volume [20]). We are counteracting this to some extent by running each task in two consecutive years.

The GREC research programme started out as an investigation into the REG task as grounded in discourse context and informed by naturally occurring data. We have since created two substantial annotated data resources which between them contain some 21,000 annotated referring expressions. These data resources, along with all automatically computed evaluation methods, become freely available for research purposes as soon as the last competition based on them is concluded.

Acknowledgments

The research reported in this paper was supported under EPSRC (UK) grants EP/E029116/1 (the Prodigy Project), EP/F059760/1 (REG Challenges 2008), and EP/G03995X/1 (Generation Challenges 2009).

References

1. Bagga, A., Baldwin, B.: Algorithms for scoring coreference chains. In: Proceedings of the Linguistic Coreference Workshop at LREC'98. pp. 563–566 (1998)
2. 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). pp. 313–320 (2006)
3. Belz, A., Varges, S.: Generation of repeated references to discourse entities. In: Proceedings of ENLG'07. pp. 9–16 (2007)

4. Belz, A.: That's nice ... what can you do with it? Computational Linguistics 35(1), 111–118 (2009)
5. Bohnet, B., Dale, R.: Viewing referring expression generation as search. In: Proceedings of the 19th International Joint Conference on Artificial Intelligence. pp. 1004–1009 (2005), http://www.ijcai.org/papers/0969.pdf
6. Charniak, E., Knight, K., Yamada, K.: Syntax-based language models for machine translation. In: Proc. MT Summit IX (2003)
7. Cheng, H., Poesio, M., Henschel, R., Mellish, C.: Corpus-based np modifier generation. In: Proceedings of NAACL 2001 (2001)
8. Choi, B., Pak, A.: A catalog of biases in questionnaires. Preventing Chronic Disease 2(1) (2005)
9. Dale, R.: Cooking up referring expressions. In: Proceedings of the 27th Annual Meeting of the Association for Computational Linguistics (1989), http://acl.ldc.upenn.edu/P/P89/P89-1009.pdf
10. Dale, R., Reiter, E.: Computational interpretations of the Gricean maxims in the generation of referring expressions. Cognitive Science 19(2), 233–263 (1995), http://www.ics.mq.edu.au/ rdale/publications/papers/1995/950426.pdf
11. van Deemter, K.: Generating referring expressions: Boolean extensions of the Incremental Algorithm. Computational Linguistics 28(1), 37–52 (2002), http://acl.ldc.upenn.edu/J/J02/J02-1003.pdf
12. van Deemter, K., van der Sluis, I., Gatt, A.: Building a semantically transparent corpus for the generation of referring expressions. In: Proceedings of the 4th International Conference on Natural Language Generation. pp. 130–132. Sydney, Australia (July 2006), http://www.aclweb.org/anthology/W/W06/W06-1420
13. Deneefe, S., Knight, K.: Synchronous tree adjoining machine translation. In: Proceedings of EMNLP (2009), http://www.isi.edu/natural-language/mt/adjoin09.pdf
14. Forster, K.I., Forster, J.C.: DMDX: A windows display program with millisecond accuracy. Behavior Research Methods, Instruments, & Computers 35(1), 116–124 (2003)
15. Gatt, A., Belz, A.: Introducing shared tasks to NLG: The TUNA shared task evaluation challenges. In: Krahmer, E., Theune, M. (eds.) Empirical Methods in Natural Language Generation, Lecture Notes in Computer Science, vol. 5980. Springer, Berlin / Heidelberg (2010)
16. Gupta, S., Stent, A.: Automatic evaluation of referring expression generation using corpora. In: Proceedings of the 1st Workshop on Using Copora in Natural Language Generation. pp. 1–6 (2005), http://www.itri.brighton.ac.uk/ucnlg/Proceedings/gupta-stent.pdf
17. Huddleston, R., Pullum, G.: The Cambridge Grammar of the English Language. Cambridge University Press (2002)
18. Jordan, P.W.: Contextual influences on attribute selection for repeated descriptions. In: van Deemter, K., Kibble, R. (eds.) Information Sharing: Reference and Presupposition in Language Generation and Interpretation. CSLI Publications, Stanford, CA (2002)
19. Jordan, P.W., Walker, M.: Learning attribute selections for non-pronominal expressions. In: Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics (2000), http://acl.ldc.upenn.edu/P/P00/P00-1024.pdf
20. 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, Lecture Notes in Computer Science, vol. 5980. Springer, Berlin / Heidelberg (2010)

21. Krahmer, E., Theune, M.: Efficient context-sensitive generation of referring expressions. In: van Deemter, K., Kibble, R. (eds.) Information Sharing: Reference and Presupposition in Language Generation and Interpretation, pp. 223–264. CSLI Publications, Stanford, CA (2002)
22. Lin, C.Y., Och, F.J.: ORANGE: A method for evaluating automatic evaluation metrics for machine translation. In: Proceedings of the 20th International Conference on Computational Linguistics (COLING'04). pp. 501–507. Geneva (2004)
23. Luo, X.: On coreference resolution performance metrics. Proc. of HLT-EMNLP pp. 25–32 (2005)
24. Morton, T.: Using Semantic Relations to Improve Information Retrieval. Ph.D. thesis, University of Pensylvania (2005)
25. Nenkova, A.: Entity-driven rewrite for multi-document summarization. In: Proceedings of IJCNLP'08 (2008)
26. Nenkova, A.: Understanding the process of multi-document summarization: content selection, rewrite and evaluation. Ph.D. thesis, Columbia University (2006)
27. Otterbacher, J., Radev, D., Luo, A.: Revisions that improve cohesion in multi-document summaries: a preliminary study. In: Proceedings of the ACL'02 Workshop on Automatic Summarization. pp. 27–36 (2002)
28. Poesio, M.: Annotating a corpus to develop and evaluate discourse entity realization algorithms: issues and preliminary results. In: Proceedings of LREC 2000 (2000)
29. Poesio, M.: Discourse annotation and semantic annotation in the GNOME corpus. In: Proc. ACL'04 Discourse Annotation Workshop (2004)
30. Qiu, L., Kan, M., Chua, T.S.: A public reference implementation of the rap anaphora resolution algorithm. In: Proceedings of LREC'04. pp. 291–294 (2004)
31. Reiter, E., Dale, R.: A fast algorithm for the generation of referring expressions. In: Proceedings of the 14th International Conference on Computational Linguistics. pp. 232–238. Nantes, France (23-28 August 1992), http://www.ics.mq.edu.au/ rdale/publications/papers/1992/C92-1038.pdf
32. Siddharthan, A., Copestake, A.: Generating referring expressions in open domains. In: Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics. Barcelona, Spain (2004), http://www.cl.cam.ac.uk/ as372/RefExp.pdf
33. Siegel, S.: Non-parametric statistics. The American Statistician 11(3), 13–19 (1957)
34. Steinberger, J., Poesio, M., Kabadjov, M., Jezek, K.: Two uses of anaphora resolution in summarization. Information Processing and Management: Special issue on Summarization 43(6), 1663–1680 (2007)
35. Trang Dang, H.: DUC 2005: Evaluation of question-focused summarization systems. In: Proceedings of the COLING-ACL'06 Workshop on Task-Focused Summarization and Question Answering. pp. 48–55. Prague (2006)
36. Turian, J., Shen, L., Melamed, I.D.: Evaluation of machine translation and its evaluation. In: Proceedings of MT Summmit IX. pp. 386–393. New Orleans (2003)
37. Viethen, J., Dale, R.: Towards the evaluation of referring expression generation. In: Proceedings of the 4th Australasian Language Technology Workshop (ALTW'06). pp. 115–122 (2006)
38. Vilain, M., Burger, J., Aberdeen, J., Connolly, D., Hirschman, L.: A model-theoretic coreference scoring scheme. Proceedings of MUC-6 pp. 45–52 (1995)