Natural Language Generation for Nature Conservation: Automating Feedback to help Volunteers identify Bumblebee Species

Steven Blake[1] *Advaith Siddharthan*[1] *Hien Nguyen*[1]
Nirwan Sharma[1] *Anne-Marie Robinson*[2] *Elaine O'Mahony*[3]
Ben Darvill[3] *Chris Mellish*[1] *René van der Wal*[2]

(1) Department of Computing Science, University of Aberdeen, U.K.
(2) Aberdeen Centre for Environmental Sustainability (ACES), University of Aberdeen, U.K.
(3) Bumblebee Conservation Trust, University of Stirling, U.K.

s.blake.08@aberdeen.ac.uk, advaith@abdn.ac.uk, h.nguyen@abdn.ac.uk,
n.sharma@abdn.ac.uk, annierobinson@abdn.ac.uk,
elaine.omahony@bumblebeeconservation.org,
ben.darvill@bumblebeeconservation.org, c.mellish@abdn.ac.uk,
r.vanderwal@abdn.ac.uk

ABSTRACT

This paper explores the use of Natural Language Generation (NLG) for facilitating the provision of feedback to citizen scientists in the context of a nature conservation programme, BEEWATCH. BEEWATCH aims to capture the distribution of bumblebees, an ecologically and economically important species group in decline, across the UK and beyond. The NLG module described here uses comparisons of visual features of bumblebee species as well as contextual information to improve the citizen scientists' identification skills and to keep them motivated. We report studies that show a positive effect of NLG feedback on accuracy of bumblebee identification and on volunteer retention, along with a positive appraisal of the generated feedback.

KEYWORDS: NLG, Natural Language Generation, Educational Application, Bumblebee Conservation, Citizen Science, Generating Feedback.

Proceedings of COLING 2012: Technical Papers, pages 311–324,
COLING 2012, Mumbai, December 2012.

1 Introduction

There is a growing realisation of the potential of digital approaches, including the use of websites and social media, to increase participation in "citizen science", which includes observing and monitoring the natural world. For instance, in the UK, the Open Air Laboratories (OPAL) network (www.opalexplorenature.org) is a large current initiative led by Imperial College, which aims to create and inspire a new generation of nature-lovers by getting people to explore, enjoy and protect their local environment (Silvertown, 2009). Within OPAL, iSpot (www.ispot.org.uk) is an online nature community that connects beginners with experts and fellow enthusiasts. Other groups have explored the use of standard social networking sites to generate public interest and collect data about the distribution of species (Stafford et al., 2010). Publicly available resources include www.thewildlab.org, which provides software for a number of mobile platforms, and www.scienceforcitizens.net, which acts as a forum for citizen scientists to find out about projects they can participate in and for researchers to publicise their projects.

Although digital tools can enthuse the public and be used to enlist (for a short time at least) willing volunteers for nature conservation projects, initiatives such as the above have to contend with at least the following issues:

1. Data quality: participants are generally untrained, and not necessarily motivated to produce high-quality data (Stafford et al., 2010). This is not a problem if a project is primarily an "outreach" activity, but many projects also have specific scientific goals.

2. Retention of volunteers: in order to secure continuing participation, systems need to constantly renew to keep users interested and engaged.

Existing digital support for nature conservation volunteers can only give them feedback and encouragement by allowing them to interact with human experts or by showing them pre-prepared material. Human experts are in short supply, and pre-prepared material is inherently limited in scope. We believe that this latter is a problem, and that the level of interest and motivation for people to participate in ecological monitoring activities is in part a function of the richness of information they are provided with on an ongoing basis. Through the use of a Natural Language Generation (NLG) component, we automate the provision of rich information to address the two key issues listed above: improving the accuracy of volunteer contributed records, and volunteer retention over time.

The Bumblebee Conservation Trust[1] is seeking to map the current distribution of bumblebee species across the UK through a collaborative project with the University of Aberdeen called BeeWatch[2]. BeeWatch allows volunteers from the general public to submit photos of bumblebees they have seen in the wild, along with the location and date of sighting. The submission interface includes an online identification guide (see Fig. 1) to help classify the bumblebee in the photo as one of 22 bumblebee species[3]. Through this interface, the volunteer can select visual features of the bumblebee (types of thorax, abdomen, etc.) to narrow down the possible species. Once the image and the user's identification have been submitted, an expert identifies

[1] http://www.bumblebeeconservation.org

[2] http://bumblebeeconservation.org/get-involved/surveys/

[3] There are actually 24 species of bumblebee in the United Kingdom, but three of these (*Bombus lucorum*, *Bombus cryptarum* and *Bombus magnus*) cannot be reliably distinguished from each other based on visual charecteristics alone. These form a species complex, and for the purposes of BeeWatch they are treated as one species— *Bombus lucorum*, the White-tailed bumblebee.

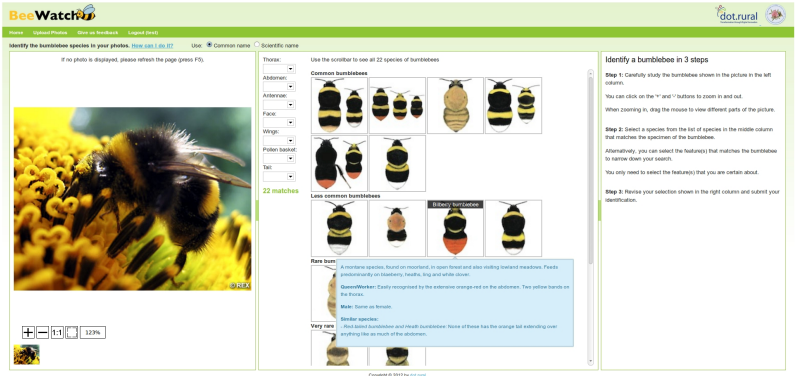


Figure 1: Screenshot of online identification tool

the bumblebee in the photo. Before the current collaboration, this expert communicated the correct identification to the volunteer by email. In the case of incorrect identification, the conservation charity would have liked to provide explanations as to why this was the case, but their experts only rarely found the time to do this. The goal of the NLG component described in this paper is to automate the provision of feedback to the volunteer based on the expert's identification.

Fig. 2 shows an extract of the computer generated feedback for a user who has incorrectly identified a photo as a broken-belted bumblebee. The key element in this automatically generated extract is the comparison between the user-identified species and the correct species as identified by the expert. Although there are only 22 species of bumblebees in the UK, this means that there are almost 500 bee comparisons to be generated, and the overall number of possible texts would be orders of magnitude greater when contextual information (e.g., based on location and time of sighting) is included in the feedback. On the other hand, the NLG

> Thank you for submitting this photo. Our expert identified the bee as a Heath bumblebee rather than a Broken-belted bumblebee. You correctly identified the face, the wings and the pollen basket; however, the abdomen (rear body) and the thorax (central body) are different. Although some of these features may not be visible in your photograph, the following advice might be helpful for next time you are in the field.
>
> The Heath bumblebee's thorax is black with two yellow to golden bands whereas the Broken-belted bumblebee's thorax is black with one yellow to golden band. The Heath bumblebee's abdomen is black with one yellow band near the top of it and a white tip whereas the Broken-belted bumblebee's abdomen is black with one yellow band around the middle of it and a white to buff tip.

Figure 2: Example of NLG feedback explaining why a user identification was incorrect

Thank you for submitting this photo. You have correctly identified the bumblebee as a White-tailed bumblebee.

As you are already aware all individuals of this species have two yellow bands and a white tail. Although they can be difficult to separate from Buff-tailed workers, the tail of the White-tailed bumblebee is pure white, with a complete absence of any buff-coloured hairs. The colour of the two yellow bands is brighter and more lemon than that seen in the Buff-tailed bumblebee.

Figure 3: Example of NLG feedback when a user has made a correct identification

system relies on just a model of each of the 22 species found in the UK. So, whereas it would be infeasible to produce a human-authored text for each possible situation, this use of NLG can even scale up to, for example, bumblebees in other countries (there are over 250 known species worldwide), or likewise, other genera.

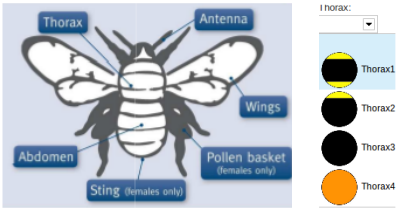
We also use feedback to reinforce relevant information when a user has made a correct identification. Fig. 3 provides an extract from computer generated feedback that illustrates this.

A key contribution of this paper is an evaluation of the effect of the automatically generated text on volunteers. It is unusual for an NLG application (that typically aid decision making in the workplace) to have as diverse a set of users as we do. Our results demonstrate the potential for NLG in real world applications targeting members of the general public. The automation of feedback provision, as described in the rest of this paper, has removed a major bottleneck for the charity we are working with, and is allowing them to scale up what was initially just a public engagement exercise generating around 200 records a year into an initiative that has produced 650 records a month.

2 Related work

Much of the recent focus within NLG applications research has been on data-to-text systems, which typically generate summaries of technical data for professionals such as engineers or nurses (Goldberg et al., 1994; Theune et al., 2001; Portet et al., 2009). These are capable of generating high quality texts; e.g., offshore oil rig workers preferred weather forecasts generated by the SumTime system to texts written by professional human forecasters (Sripada et al., 2003). There is some previous work on the use of data-to-text for lay audiences; e.g., generating narratives from sensor data for automotive (Reddington et al., 2011) and environmental (Molina et al., 2011) applications, generating personal narratives to help children with complex communication needs (Black et al., 2010), and summarising neonatal intensive care data for parents of premature babies (Mahamood et al., 2008).

There are some notable examples of NLG systems that make use of structured textual records rather than numeric data. Peba-II (Milosavljevic, 1997) was an online animal encyclopedia that provided descriptions and comparison of animals using HTML pages. The Power (Daley et al., 1998) and ILEX (O'Donnell et al., 2001) systems in the virtual museum domain dynamically generated descriptions of museum objects based on the user's discourse history and user model. Dial Your Disc (Van Deemter and Odijk, 1997) generated spoken monologues about classical music, with the aim of generating engaging texts, attempting to keep its users amused by



(a) (b)

Figure 4: Bumblebee Model: (a) Visual bumblebee features, (b) List of Thorax types

focusing on the expression of *unusual* content. Our work shares commonalities with these systems, in that it targets non-expert audiences and has educational goals.

The main mechanism employed by us in the current work is the use of comparisons for generating feedback. This builds on a body of previous work. Milosavljevic (1997) described comparisons as a useful tool for augmenting the user's existing knowledge with new knowledge. Karasimos and Isard (2004) showed that texts which contained comparisons and aggregations helped readers retain more information and perform better on factual recall while also finding these texts more interesting and pleasant to read. Later, Marge et al. (2008) performed a similar experiment to isolate the effects of comparison from those of aggregation. These two experiments provide evidence that comparisons can help to improve the knowledge of users on a given domain, and are a basis for our work.

3 Implementation of the NLG module

Our NLG system uses the architecture proposed by Reiter and Dale (2000) and is compatible with a wide range of work within the field. There are three main components in this architecture; a document planner, a micro planner and a surface realiser. Additionally, the document planning makes use of a domain model. We describe our implementation in this section and the evaluation of the system in Sections 4 and 5.

3.1 Domain model

Our domain model is primarily the representation of each bumblebee species as a set of visual features, as implemented in the online identification guide (see Fig. 1). The tool contains information about the thorax and abdomen colour patterns, face length, wing type and the presence of a pollen basket (shown in Fig. 4 (a)). The model includes a textual description of each visual feature. For example, the thorax with visual pattern `Thorax2` in Fig. 4 (b) has the associated text: "black with one yellow to golden band".

Note that the domain model used in the identification tool only covers the main bumblebee features. This is important for usability, and also makes it easier to port to other insect species. However, this means that some bumblebee species are indistinguishable in terms of the modelled features. To distinguish such species, each model of a bumblebee species has associated with it a list of similar bumblebee species and a description of how it can be distinguished from them. An example of this similarity can be seen between the Moss carder bee and the Common carder bee (shown in Fig. 5(a)). They share the same values for all their domain features (e.g., both have thorax type 4, etc.), and their only distinguishing feature is the black hairs on the

(a) (b)

Figure 5: Difficult cases: (a) Similarity of Moss Carder and Common Carder Bumblebee, (b) Differences between castes for the White-tailed bumblebee

abdomen of the Common carder bee. The model also contains contextual information; for instance, where the species is usually found, what time of year they are usually seen and how rare they are in the UK.

A further point worth noting is that bumblebees have a caste system of queens, males (drones) and females (workers). These castes can sometimes have different visual features (e.g., see Fig. 5(b)). We do not model castes explicitly because it is difficult for novice recorders to identify caste, and there is thus limited gain in using caste information in the feedback. As an alternative, we allow features for the abdomen and thorax to sometimes have multiple values for the same species.

3.2 Document planner

The NLG system described here primarily uses comparisons of bumblebee species to improve the user's identification skills. Milosavljevic (1997) introduced three types of comparison that are useful for educational purposes: direct, clarifactory and illustrative. Briefly, direct comparison is used to describe to the user attributes that two objects share and the attributes that distinguish them. This comparison is bi-focal; neither object is more important than the other. Clarifactory comparison is used to describe an object by distinguishing it from a similar object that it may be confused with. This confusion usually arises from the two objects sharing similar attributes or features. Illustrative comparison is used to describe an attribute (or attributes) of an object by referring to the same attribute(s) of another object that the user is familiar with.

The main methods used in this work are direct comparisons, which allows the user to understand the differences between species, and clarifactory comparisons to distinguish similar species (those with identical features in our visual model). Illustrative comparison is currently not used but could be employed when the system has access to a user's history of interaction with the system. For example if the user had identified a Garden bumblebee previously and has just identified a Heath bumblebee, illustrative comparison could be used to explain that the colour pattern of the Garden bumblebee is similar to that of the Heath bumblebee. However, users typically submit records only once or twice a week, so past submissions might not be salient.

Our document planner first determines content, and then decides document structure.

3.2.1 Content determination

Before the system can decide what to include it needs a data representation to store this content. This representation will, later on, map to a sentence (or a group of related sentences) that will be generated in the final output. We represent content using *messages*, i.e. Java objects that contain domain specific entities and information.

The creation of message objects is handled by the message generator at run-time. The message generator looks at the input the system has been given and then analyses it to determine which of a set of content determination rules match. These rules determine which messages are created and what they contain.

A `thankYou` message is always generated with canned text content thanking the user for the submission. Then a message is generated that holds information about whether the user has identified the bee correctly or not. For example, if the identification by the user is incorrect then a `result` message is generated containing the value "incorrect" along with the bumblebee species as identified by the expert and by the user. Further along the line, this message will be realised as a sentence such as "*Our expert identified the bee as a Heath bumblebee rather than a Broken-belted bumblebee*" (see Fig. 2). If the identification is correct, a similar message is generated with the value "correct", which could later be realised as a sentences such as "*You have correctly identified the bee as a Heath bumblebee*".

If the user has incorrectly identified the bumblebee species, the system needs to explain to the user why the identification was incorrect. This is handled through the `features` message. This message compares the visual features of the two species and stores similar features and dissimilar features as two separate sets. Due to the fact that some features (e.g., thorax and abdomen) can have multiple values in a bumblebee model due to variation between castes, the similarity of features is determined through set intersection. A feature goes into the "similar" list, if there is any value that is common to both bumblebee models. This message is later realised as a sentence such as "*You correctly identified the face, the wings and the pollen basket; however, the abdomen (rear body) and the thorax (central body) are different.*"

One of our communicative goals is to help the user improve their identification skills. This is pursued by explaining to the user why they were incorrect (if that were indeed the case). For each feature that is dissimilar, a `featureIdentification` message is created that contains the values of this feature in both bumblebee models. These messages will be used to explain the differences between the two species using direct comparison.

The message generator also searches the "similar species" list in the model of the bumblebee as identified by the expert. If the bumblebee as identified by the user is found, then a `similarSpecies` message is created, containing the canned text from the model describing the distinguishing features using clarifactory comparison.

If there is any contextual information available about the identified species, this is realised through a `contextualInformation` message. In our current implementation, we only use a summary of the known habitat and behaviour of the species in a non-adaptive manner. We plan to use such information more intelligently in the future, notably to help contextualise a user's record.

When all the messages have been constructed, they are passed on to the Document Structuring component.

3.2.2 Document structuring

The messages are structured in a specific way by the document planner using schemas, resulting in a document plan which takes the form of a tree. Groups of related messages are represented by the internal nodes while the messages themselves are the terminal nodes. In this case, the internal nodes will represent actual textual structures such as paragraphs. These structures are

represented in Java as objects that are instantiated from classes. The design of these structures is heavily influenced by the structures proposed by Reiter and Dale (2000).

The exact structure is determined by how many messages have been generated. There are four possible groups that the messages can be placed in: the intro group, the features group, the similar species group and the contextual information group. The intro group will always be present but the other three may or may not be present depending on the outcome of the identification and the two species that are being compared. For example, if two species are not classed as similar species then the similar species group will not be present at all. The schema loops through the entire message list and decides what groups are needed based on the type of messages present.

3.3 Microplanner

The microplanner is the second stage of the generation architecture. Here, the document plan produced by the first stage, the document planner, is refined to produce a text specification. This includes phrase specifications and their aggregation into sentences. The phrase specification is structured in a way that it can be realised by the surface realiser. We use SIMPLENLG (Gatt and Reiter, 2009), and the design of these structures is therefore influenced by the functionalities of the SIMPLENLG library.

To allow for the generation of more complex sentence structures and sentences that are easier to understand, the microplanner also carries out aggregation. The aggregation performed in this system focuses on the formation of sentences. For example, in the paragraph that deals with comparisons (see Fig. 2), the system knows that it will be comparing different features. This means that the phrase specifications can be aggregated through subordination, using conjunctions such as “whereas” or “although”.

The descriptions of the visual features, specifically the abdomen and thorax, are also analysed to check if they are a candidate for aggregation. These two features can have multiple values due to morphological differences among castes within a species. Our system does not explicitly model castes and therefore when describing a feature that has multiple values the system has to use disjunctions. For example, the White-tailed bumblebee has two possible thorax values: `thorax1` (“black with two yellow to golden bands”) and `thorax2` (“black with one yellow to golden band”; see Fig. 5). Rather than generating:

> “the thorax is either black with two yellow to golden bands or black with one yellow to golden band,”

we use aggregation to generate the more natural sounding succinct phrase:

> “the thorax is black with either one or two yellow to golden bands.”

Once the microplanner has built all the sentence specifications (by the processes mentioned previously) and organised them into paragraphs, the resultant text specification document is passed on to the surface realiser.

3.4 Surface realiser

The role of the surface realiser is to convert the text specification received from the microplanner into text that the user can read and understand. This includes linguistic realisation (converting

the sentence specifications into sentences) and structural realisation (structuring the sentences inside the document). Both the linguistic and structural realisations are performed by using functionalities provided by the SIMPLENLG realiser library (Gatt and Reiter, 2009).

4 Experiment 1

This section details the evaluation of how volunteer recorders perceive the NLG, and the effect of feedback on their bumblebee identification accuracy. Later, in Experiment 2 (Section 5), we study the effect of NLG feedback on volunteer retention. For Experiment 1, we distinguish between two types of feedback, which differ with respect to the richness of information they communicate:

Type 1: Acknowledgement of submission + Correct Answer

Type 2: Type 1 + Feedback based on comparisons of visual features

4.1 Method

We designed an evaluation interface that steps each participant through twenty distinct images of bumblebees for which we have an expert identification. There were seven species of bumblebee represented in this data set, each occurring between one and four times. At each step, the participant identified the bumblebee species in the photo using an identification guide, and then received feedback on their identification. All participants viewed the photos in the same order. Upon completing the identification task for all 20 photos, each user was asked:

- To rate how helpful they found the feedback (on a scale of 1–5)
- What they thought of the exercise in general (free text)
- What extra types of information they would like to see in the feedback (free text)?
- What information they would like to see removed (free text)

To test the effect of feedback type on recording accuracy, we randomly divided our participants into two groups. Group A always received Type 1 feedback, while Group B always received the richer Type 2 feedback[4]. 48 participants completed Experiment 1, 21 in Group A and 27 in Group B. All participants were undergraduate students studying biology at the University of Aberdeen and none had prior experience with bumblebee identification. The use of naive participants was deliberate. Most of the volunteers to the BEEWATCH program start out as naive participants, and the role of training through feedback about identification features is likely to be most useful at this stage.

4.2 Results

4.2.1 Effect of feedback on accuracy

The accuracy of bumblebee identification for both groups was analysed. We expected that initially there would be no difference between the two groups, but that over time, Group B would improve their identification accuracy faster than Group A due to the more informative

[4]There is a further Type 3 feedback that our system generates, which includes contextual information about the bumblebee species. However this is mainly useful for identification in the real world, and we did not consider it for Experiment 1. We use Type 3 feedback later in Experiment 2, which was conducted with volunteers submitting photos to the live BEEWATCH website

Identifications	Acc(Type2) - Acc(Type1)	Significance of Difference
images 1-10	-2.7%	p=.55
images 11-20	7.4%	p=.10

Table 1: Accuracy for identifications 1-10 and 11-20

feedback received. Table 1 shows that the difference in accuracy between the two groups for the first ten identifications is minimal. For the last ten identifications, however, there is a bigger difference between the means of the two groups, and the group receiving NLG feedback is performing better by 7.4% points.

To understand the effect of Type 2 feedback in more detail, we performed a generalised linear mixed model fit by the Laplace approximation. The dependent variable was Accuracy (whether a photo was identified correctly). The independent variables were Time (the order of presentation of photos) and Condition (Type 1 or Type 2 feedback). We expected that there would be differences between participants and that some bee species would be easier to identify than others. To accommodate these expectations, Participant and BeeSpecies were included as random factors in the model.

We found a significant main effect of Time (z=-1.768; p=0.0423) and Condition (z=2.03l; p=0.017); i.e, both groups improved over time, and Group B was overall more accurate. More importantly, we also found a strong interaction between Time and Condition (z=-3.260; p=0.001); i.e., the accuracy of Group B increased faster over time than Group A. Thus, the richer Type 2 feedback proved useful for improving recorder skills over time.

The graph of the difference in accuracy between the two groups over time in Fig. 6 illustrates this finding. Positive values indicate that the mean accuracy of Group B is greater while negative values indicate that the mean accuracy of Group A is greater. The mix of positive and negative values for the first fifteen identifications show that neither of the groups are consistently more accurate than the other. However, the last five identifications show continually positive values, indicating that the richer feedback received by Group B was beginning to take effect.

These results are consistent with those reported in Karasimos and Isard (2004) and Marge et al. (2008), which collectively suggest that the use of comparison helps users to retain information.

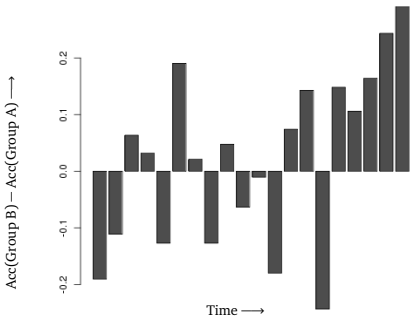


Figure 6: Graph of difference in accuracy between groups over time

Type 1 (no NLG)	Type 2 (NLG)
3.09	3.85

Table 2: Helpfulness of feedback: Mean score for each condition

4.2.2 Feedback helpfulness

One of the questions that participants were asked was to rate how helpful they found the feedback on a scale of 1–5. The NLG feedback was perceived to be significantly more helpful (t-test; t=2.78; p <0.01), as shown in Table 2.

The 21 participants in Group A who received Type 1 feedback overwhelmingly requested more feedback in their free text answers:

- Would like to have been given information on why their identification was incorrect (14 participants)
- Would like none of the information to be removed (12 participants)
- Would like to see more information (3 participants)
- Thought there was a lack of information (3 participants)
- Would like to have been told some facts about the bumblebee species that had been identified (2 participants)

From the feedback listed above, it is clear that participants would like to see more information than just a "correct/incorrect" response; specifically, explanation as to why their identification was incorrect was requested my most participants. There was also a small proportion who would like to be presented with contextual information about the bumblebee identified.

The 27 participants in Group B who received Type 2 feedback were more satisfied with the richness of information in the feedback:

- Would like none of the information to be removed (11 participants)
- Thought that no more information should be added and that the level of detail was sufficient (8 participants)
- Found the information to be useful (5 participants)
- Would like to see comparative pictures between the two bumblebee species (2 participants)
- Some comparisons didn't go into a lot of detail, specifically between similar species (2 participants)
- Face shape was rarely visible so did not help in identification (2 participants)
- Would like to see some contextual information about the bumblebee species (rarity, life history etc) (1 participant)

It is clear from the user feedback that users were more or less satisfied with Type 2 feedback, and that Type 1 users would have preferred Type 2 feedback. Further, a small proportion from both groups wanted more contextual information. As we report next, the inclusion of such information in Type 3 feedback has a positive effect on volunteers.

5 Experiment 2

The experiment described above showed that the richer feedback provided by our NLG component helped improve recorder accuracy. The other key issue we are interested in is volunteer motivation. We present an initial evaluation of this using our live system, by comparing the number of submissions made by users who receive three different types of feedback (volunteer return rates).

Feedback	No. of Submissions	No. of Users	Submissions/User	χ^2
Type 1	356	110	3.23	
Type 2	412	123	3.35	$p = 0.04$
Type 3	542	104	5.21	$p < 0.0001$

Table 3: Number of submissions by feedback type

5.1 Method

The NLG component described in this paper is being used live by the Bumblebee Conservation Trust since June 26 2012. When a photograph and volunteer identification of the bumblebee therein are submitted, the information is logged in a database. Periodically an expert reviews the database using a special administrator tool. When an expert calls up a particular submission and enters their own identification of the bumblebee in the photo, the NLG text is automatically provided. The expert can then edit this text if necessary (to address a question asked by the recorder, for instance) and then clicks on a button to send the feedback by email. This interface has dramatically increased the throughput of the experts, enabled them to spend more time on difficult cases, and giving the charity the confidence to publicise the project in the media and scale it up in size. The interface also means that the expert feedback reaches volunteers quicker, which presumably helps to motivate users and encourage further submissions.

When users register for the live system, they are randomly allocated to one of the three groups, by dividing their unique user IDs by 3 and using the remainder. We thus expect similar numbers of users in each group. As not everyone who registers submits records, group sizes are not identical (see Table 3). Each group always received feedback in only one of the following three types:

Type 1: Acknowledgement of submission + Correct Answer

Type 2: Type 1 + Feedback based on comparisons of visual features

Type 3: Type 2 + Contextual information

For the experiment, the administration interface made the experts aware of the three feedback conditions and which users were in which group. The automatic NLG feedback reflected the feedback condition and the experts were instructed to only make edits compatible with the prevailing condition.

5.2 Results

We report preliminary results from the first 2 months of the system going live. We received 1310 submissions from 337 participants during this period. Table 3 shows the number of submissions by experimental condition. There was a significant difference in submission numbers ($\chi^2 = 41.338$; 2 degrees of freedom; $p < 0.0001$) for the three treatment groups. The table shows the pair-wise significance for differences of Type 2 and 3 from Type 1 feedback.

As users who received the richer Type 2 or Type 3 feedback submitted more records on average than users who received Type 1 feedback, it appears that increasing the richness of information provision through NLG feedback has a positive effect on return rates of participants to the website. However, this analysis is preliminary and these figures are potentially skewed by the presence of a small number of dedicated volunteers in each feedback group.

On average, users submitted 3.7 photos during this period. This is insufficient to test improvement in accuracy from the live system at this point in time.

Conclusion and perspectives

We have described an NLG system that is being used by a nature conservation charity for a citizen science initiative. The automation of feedback provision has removed a major bottleneck for the charity, and has allowed them to scale up what was initially just a public engagement exercise generating around 200 records a year into an initiative that has produced 650 records a month. We are also investigating crowd sourcing models to reduce the time commitments on experts even further and allow a further scaling up of the initiative. These models work better with more accurate identifications by individuals; thus improving recorder accuracy is vital.

Our results show that the feedback generated by the NLG system described in this paper helps users to improve their identification skills faster that those who only receive the correct answer as feedback. Users also found the feedback produced through NLG more helpful than the simple feedback with no NLG elements, as evident from both qualitative and quantitative data reported for Experiment 1, and more motivating, as evident from the return rates reported for Experiment 2.

Acknowledgments

This research was supported by an award made by the RCUK Digital Economy programme to the dot.rural Digital Economy Hub; award reference: EP/G066051/1. We would also like to thank the Bumblebee Conservation Trust for making available their resources, including the use of their organisational infrastructure to recruit volunteers, the provision of experts to identify submitted records, and their contributions towards the testing of the NLG component. Further, this research has depended on records submitted by volunteer recorders participating in the BeeWatch programme.

References

Black, R., Reddington, J., Reiter, E., Tintarev, N., and Waller, A. (2010). Using nlg and sensors to support personal narrative for children with complex communication needs. In *Proceedings of the NAACL HLT 2010 Workshop on Speech and Language Processing for Assistive Technologies*, pages 1–9. Association for Computational Linguistics.

Daley, R., Greeny, S. J., Milosavljevicz, M., Parisz, C., Verspoory, C., and Williamsy, S. (1998). The realities of generating natural language from databases. In *Proceedings of the 11th Australian Joint Conference on Artificial Intelligence*.

Gatt, A. and Reiter, E. (2009). Simplenlg: A realisation engine for practical applications. In *Proceedings of ENLG-2009*.

Goldberg, E., Driedger, N., and Kittredge, R. (1994). Using natural-language processing to produce weather forecasts. *IEEE Expert*, 9(2):45–53.

Karasimos, A. and Isard, A. (2004). Multi-lingual evaluation of a natural language generation system. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation*.

Mahamood, S., Reiter, E., and Mellish, C. (2008). Neonatal intensive care information for parents an affective approach. In *Computer-Based Medical Systems, 2008. CBMS'08. 21st IEEE International Symposium on*, pages 461–463. IEEE.

Marge, M., Isard, A., and Moore, J. (2008). Creation of a new domain and evaluation of comparison generation in a natural language generation system. In *Proceedings of the Fifth International Language Generation Conference*.

Milosavljevic, M. (1997). Augmenting the user's knowledge via comparison. In *In Proceedings of the 6th International Conference on User Modelling*.

Molina, M., Stent, A., and Parodi, E. (2011). Generating automated news to explain the meaning of sensor data. *Advances in Intelligent Data Analysis X*, pages 282–293.

O'Donnell, M., Mellish, C., Oberlander, J., and Knott, A. (2001). Ilex: An architecture for a dynamic hypertext generation system. *Natural Language Engineering 7*.

Portet, F., Reiter, E., Gatt, A., Hunter, J., Sripada, S., Freer, Y., and Sykes, C. (2009). Automatic generation of textual summaries from neonatal intensive care data. *Artificial Intelligence*, 173(7-8):789–816.

Reddington, J., Reiter, E., Tintarev, N., Black, R., and Waller, A. (2011). "Hands Busy, Eyes Busy": Generating Stories from Sensor Data for Automotive applications. In *Proceedings of IUI Workshop on Multimodal Interfaces for Automotive Applications*.

Reiter, E. and Dale, R. (2000). *Building Natural Language Generation Systems*. Cambridge University Press, Cambridge, UK.

Silvertown, J. (2009). A new dawn for citizen science. *Trends in Ecology & Evolution*, 24(9):467–471.

Sripada, S., Reiter, E., and Davy, I. (2003). SumTime-Mousam: Configurable marine weather forecast generator. *Expert Update*, 6(3):4–10.

Stafford, R., Hart, A., Collins, L., Kirkhope, C., Williams, R., Rees, S., Lloyd, J., and Goodenough, A. (2010). Eu-Social Science: The Role of Internet Social Networks in the Collection of Bee Biodiversity Data. *PloS one*, 5(12):e14381.

Theune, M., Klabbers, E., de Pijper, J., Krahmer, E., and Odijk, J. (2001). From data to speech: a general approach. *Natural Language Engineering*, 7(01):47–86.

Van Deemter, K. and Odijk, J. (1997). Context modeling and the generation of spoken discourse. *Speech Communication*, 21(1-2):101–121.