Multi-lingual Evaluation of a Natural Language Generation System

Athanasios Karasimos*, **Amy Isard**†

*Theoretical and Applied Linguistics, University of Edinburgh
40 George Square, Edinburgh, UK
†School of Informatics, University of Edinburgh
2 Buccleuch Place, Edinburgh, UK
{alderk@yahoo.co.uk, amy.isard@ed.ac.uk}

Abstract

This paper describes a user evaluation of the text output from the M-PIRO (Multilingual Personalised Information Objects) system, which dynamically generates descriptions of exhibits for a virtual museum. We show that subjects performed significantly better in a factual recall test when the descriptions included more sophisticated text structuring modules. The subjects also judged the structured texts to be more interesting and readable, and felt that they had learned more from them.

1. Introduction

Many natural language generation systems use text structuring components to perform enhancements such as aggregation or comparisons, but very few formal evaluations have been done to confirm that users actually appreciate them and benefit from them in terms of understanding and information retention. This study attempts to make a first step toward providing such an evaluation for the M-PIRO (Multilingual Personalised Information Objects) system (Isard et al., 2003), which dynamically generates descriptions of exhibits for a virtual museum.

First, a brief description of the generation system will be given, focusing on the modules which implement aggregation and comparisons, and then the evaluation procedure and results will be discussed.

2. The M-PIRO System and Text Enhancements

The M-PIRO generation system was designed to be used as part of a virtual museum, either on the web, or in a virtual or augmented reality installation. It generates descriptions, which are tailored to the individual user, in one of three languages: English, Greek and Italian. Its architecture was based on the earlier ILEX (Intelligent Labelling Explorer) system (O'Donnell et al., 2001) which produced labels in English for exhibits in a museum jewelery gallery. Both systems use a typical generation architecture composed of four stages: content selection, text planning, microplanning and surface realisation (Reiter and Dale, 2000).

The information about museum exhibits is stored in a hierarchical database which provides specifications for the basic classes of the entities present in the domain (exhibit, historical period), as well as the entities themselves (exhibit1, Apollo, the Hellenistic period). As a user navigates through the museum, the system keeps a record of the exhibits viewed and avoids repeating information where possible. The system also has different settings for three user types: adult, child and expert, and facts in the database are assigned scores according to how important and interesting they are considered to be for each type by the expert curator involved in inputting the information.

2.1. Aggregation

The Content Selection module packages information into verb-based, clause-sized propositions. Texts composed exclusively of sentences based on such single-fact propositions are very likely to contain repetitions and redundancies, and are almost certain to be considered boring and unnatural by human readers. To overcome this problem, the M-PIRO system (in common with many such systems) makes use of aggregation algorithms which combine semantically related propositions in order to produce a more concise and coherent text (Melengoglou, 2002).

For example, given the three sentences:

> *This exhibit is a lekythos. It was painted with the black figure technique. It originates from Attica.*

the aggregation module could produce

> *This exhibit is a lekythos; it was painted with the black figure technique and it originates from Attica.*

The underlying decisions on, for example, when a semicolon is appropriate, and which propositions can actually be combined into an aggregated sentence, depend on a range of factors including the type of proposition and the user type (children will typically be presented with fewer aggregated structures).

2.2. Comparisons

The comparison module takes advantage of the hierarchical nature of the domain information in the M-PIRO database by grouping entities in the database in terms of common attributes (Lisowska, 2002; Melengoglou, 2002). For example, as a user navigates through the virtual museum, they may come across several exhibits of the same type, in which case the system can generate an introduction such as

> *This exhibit is another lekythos. Like the previous lekythos, it originates from Attica.*

It is also possible to create negative comparisons such as

> *Unlike the previous vessels, which were created during the classical period, this amphora was created during the archaic period.*

In these cases the system is again using fairly sophisticated techniques to identify which features the previous exhibits have in common, and which are contrastable with the new description.

3. Method

3.1. Experimental Design

Our hypotheses were that texts which contain aggregation and comparisons as described above would help readers to retain more information and perform better on a factual recall text, and that readers would rate these texts as more interesting and pleasant to read. The structure of the experiment was informed by an evaluation carried out on the ILEX system in which learning outcomes of subjects who used the dynamic hypertext version of the system were compared to those who used a static hypertext version (Cox et al., 1999)

The evaluation was carried out on both English and Greek texts. Two sequences of six exhibits each were chosen for the test suite; one concerned ancient Greek coins, and the other ancient Greek vessels. For each sequence two versions were prepared, the only difference between them being that one used the aggregation and comparison modules and the other did not. All the texts were created with the adult user type setting.

The subjects tested were 20 adult Greek speakers and 20 adult English speakers who did not have expert knowledge of Ancient Greek archaeological objects. For each language the subjects were assigned at random to one of two equally sized groups: Group A who read the texts about vessels in the structured version and coins in the plain version, and Group B who read the coins in structured version and the vessels in the plain one. After each set of texts, the subject was asked to complete a multiple-choice test and after they had completed the whole experiment, a usability questionnaire was given to them.

A pilot study (see Karasimos (2003)) had shown that subjects considered the texts which described ancient Greek vessels to be significantly more difficult to understand than those which described coins, and this was subsequently borne out by the multiple choice scores of the main experiment and the usability questionnaire (see section 4.2.). Therefore it was decided to present the coins texts first in all cases. We considered that the possibility of an ordering effect was less serious than the problem of subjects failing to concentrate because of tiredness after reading the "more difficult" set of texts.

3.2. Example Texts

The examples below, taken from the texts presented to the subjects, illustrate the differences between the plain and structured texts. At this point, the subject has already read descriptions of several ancient Greek vessels.

The following is is the first paragraph of the plain text description of a stamnos, where no reference is made to any previous objects, and each piece of information is conveyed in a separate sentence.

> *This exhibit is a stamnos. It was created during the classical period. It dates from circa 420 B.C. It has a picture of Dionysus (centre) being garlanded by maenads in a state of ecstasy. One maenad (left) is filling a skyphos with wine, another (right) is playing a drum. This stamnos was painted by the painter of Dinos. It was decorated with the red figure technique. It is made of clay.*

The next example is the equivalent structured text, where several sentences are aggregated, and comparisons are made with previous objects.

> *This exhibit is a stamnos. Unlike the previous vessels, which were created during the archaic period, this stamnos was created during the classical period. It shows Dionysus (centre) being garlanded by maenads in a state of ecstasy. One maenad (left) is filling a skyphos with wine, another (right) is playing a drum. This stamnos was decorated by the painter of Dinos with the red figure technique and is made of clay.*

Below is a similar example in Greek, taken from first the plain text and then the structured text descriptions of a tetradrachm (a type of coin).

> Αυτό το έκθεμα είναι ένα τετράδραχμο. Δημιουργήθηκε κατά τη διάρκεια της ελληνιστικής περιόδου. Χρονολογείται στον 2ο αιώνα π.Χ. Απεικονίζει μια ασπίδα και στη μέση μια προτομή, όπως συνηθιζόταν στα μακεδονικά νομίσματα. Αυτό το τετράδραχμο έχει φτιαχτεί από ασήμι.

> Αυτό το έκθεμα είναι άλλο ένα τετράδραχμο, που δημιουργήθηκε κατά τη διάρκεια της ελληνιστικής περιόδου. Χρονολογείται στον 2ο αιώνα π.Χ. Απεικονίζει μια ασπίδα και στη μέση μια προτομή, όπως συνηθιζόταν στα μακεδονικά νομίσματα. Όπως τα προηγούμενα νομίσματα, αυτό το τετράδραχμο έχει φτιαχτεί από αυτό ασήμι.

3.3. Questionnaires

The multiple choice test consisted of 15 questions, and the highest possible score was 17. Some of the multiple choice questions concerned just one exhibit, for example:

> *8. Which picture does the stamnos exhibit show?*
> *a. A marriage feast*
> *b. A man preparing to throw the javelin*
> *c. Dionysus surrounded by maenads*
> *d. A young man sitting down and writing with a stylus*

and others required the subjects to consider the texts which they had read as a whole:

14. What is the characteristic which the fewest of these exhibits have in common?
a. the creation period
b. the painting technique
c. original location
d. museum location

The usability questionnaire asked for subjective opinions about the quality and interest of the texts. The subjects were asked to assign numerical scores to quantify their interest in and enjoyment of the texts and to estimate how much they learned from them. They were also asked to give a score for how difficult they found the questions. In addition they were asked to state which of the two sets of texts they preferred from a point of view of fluency.

The full set of texts and questions can be found in Karasimos (2003).

4. Results

The quantitative scores from the multiple choice test will be discussed first and then the results of the qualitative usability questionnaire.

4.1. Multiple Choice Scores

			Structured	Plain
E	A	mean (stdev)	**13.7** (1.95)	11 (3.24)
N	B	mean (stdev)	**12.4** (2.80)	9.4 (2.80)
	A+B	mean (stdev)	**13.05** (2.44)	10.4 (3.12)
G	A	mean (stdev)	**12.2** (2.05)	11.1 (1.73)
R	B	mean (stdev)	**12.9** (2.56)	10 (2.16)
	A+B	mean (stdev)	**12.55** (2.28)	10.55 (1.99)
All		mean (stdev)	**12.8** (2.34)	10.5 (2.58)

Table 1: Summary of Results (Group A read structured vessels texts, Group B read structured coins texts)

A summary of the results from the multiple choice questionnaires is presented in table 1. The average scores were higher for the structured texts in all cases.

Graphs of the scores for individual subjects appear in figure 1 (English) and figure 2 (Greek). Only one subject performed better on the texts they read in the plain version and many performed much better on the structured version.

Our hypothesis was that subjects would learn more from reading structured texts than from reading plain ones. We performed a separate two-way repeated measures ANOVA test on the multiple choice test results for each language with the text type (simple or structured) as within-subject factor. The subjects' group (whether they saw the coins or vessels texts in the structured version) was included as a between-subjects factor.

When the English subjects were considered alone, the text type factor was significant ($F(1,18) = 39.44$, $p < .001$) and there was no significant interaction with the group factor.

The text type factor was also significant for the Greek subjects ($F(1,18) = 48.32$, $p < .001$). In this case there was a small but significant interaction with the group factor

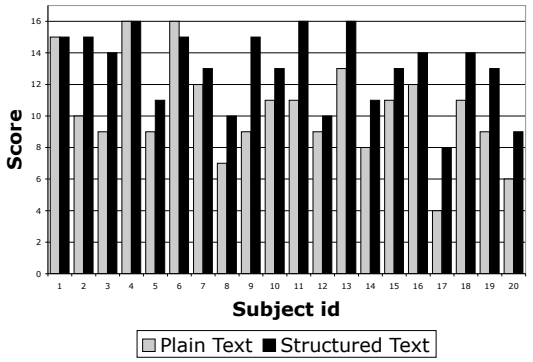

Figure 1: English Subjects

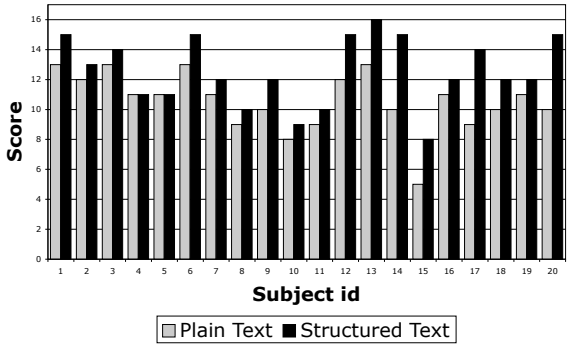

Figure 2: Greek Subjects

($F(1,18) = 9.79$, $p < .01$). In both groups, the subjects performed better with the structured text, but there was a much greater variability in the group who read the coins texts in the structured version. A discussion of some reasons for differences between the results for the two language groups is in section 5.

We also performed two-way repeated measures ANOVA test with the text type as within-subject factor and two between-subjects factors: group and language. The text type factor was again significant ($F(1,36) = 82.90$, $p < .001$), and there was no significant interaction with either the group or the language factor.

These results therefore support our hypothesis that the subjects learned more from the texts which they read in a structured version.

4.2. Usability Questionnaires

The results from the usability questionnaire (see table 2 show that the subjects considered the structured texts to be more fluent and interesting than the others, and thought that they had learned more from them.

Most of the usability questionnaire results showed trends towards a preference for the structured texts but were not statistically significant, however the when the subjects were asked to rate how much they had learned from each text on a scale of 1 to 5 (see table 3, there was a signifi-

	Structured	No Pref	Plain
more fluent	**75%**	12.5%	12.5%
more educational	**50%**	37.5%	12.5%
more interesting	35%	**40%**	25%
more enjoyable	15%	**75%**	10%

Table 2: Subjects' Preferences by Type of Text

cant difference between the mean scores ($F(1,36) = 18.79$, $p < .001$). This shows that the subjects were aware that they had retained more information from the structured texts, even though at this point they had not yet been given their scores from the multiple choice test.

	Structured	Plain
mean	**3.65**	3.12
stdev	.98	1.20

Table 3: Subjects' Estimation of Learning

For a more detailed discussion of the experimental design, pilot studies, and results, see (Karasimos, 2003).

5. Discussion and Conclusions

This study shows that the use of text structuring techniques in the M-PIRO system not only results in descriptions which are more pleasant to read, but also has a significant positive effect on the quality of the texts from an educational point of view. This is an important first step in the evaluation of a generation system.

One aspect of the study which was not fully explored was the difference in prior experience of the subject matter between the English and Greek subjects. The names of the exhibits would have been familiar to a native Greek speaker, but most would have been completely new to a native speaker of English (e.g. "kylix", "lekythos"). In addition, the Greek speakers had greater background knowledge of the historical periods from which the exhibits date. This seems to have had both positive and negative effects on performance, for example causing subjects to make assumptions about the provenance of coins based on prior knowledge rather than relying entirely on facts which they had read during the experiment. However, we consider that this did not have an effect on the results of the experiment, as we were testing the difference in results obtained from reading more or less structured texts, rather than absolute scores.

Several subjects performed very well on the multiple choice test, with near perfect scores from both sets of texts, so there was a possible ceiling effect with the results, as there was no scope for them to do any better on one set than the other. It would therefore be useful in a future study to attempt to add some more difficult questions to make this effect less likely.

In the future, we would like to compare texts produced by the natural language generation system with those written by a human curator. The handwritten texts would not contain the comparisons which our system generates, but we would expect subjects to consider them more fluent, and we would like to study the interaction between these two effects to see whether the experience of visitors to a virtual museum is enhanced by the text structuring of which our system is capable.

6. References

Richard Cox, Mick O'Donnell, and Jon Oberlander. 1999. Dynamic versus static hypermedia in museum education: an evaluation of ILEX, the intelligent labelling explorer. In *Proceedings of the Artificial Intelligence in Education conference (AI-ED99)*, Le Mans, July.

Amy Isard, Jon Oberlander, Ion Androutsopoulos, and Colin Matheson. 2003. Speaking the users' languages. *IEEE Intelligent Systems*, 18(1):40–45. Special Issue on Advances in Natural Language Processing.

Athanasios Karasimos. 2003. Evaluation of the M-PIRO text output. Master's thesis, University of Edinburgh.

Agnes Lisowska. 2002. The design and implementation of an architecture for using comparisons in the M-PIRO domain. Master's thesis, University of Edinburgh.

Alexander Melengoglou. 2002. Multilingual aggregation in the M-PIRO system. Master's thesis, University of Edinburgh.

Mick O'Donnell, Chris Mellish, Jon Oberlander, and Alistair Knott. 2001. ILEX: An architecture for a dynamic hypertext generation system. *Natural Language Engineering*, 7:225–250.

Ehud Reiter and Robert Dale. 2000. *Building Natural Language Generation Systems*. Cambridge University Press.