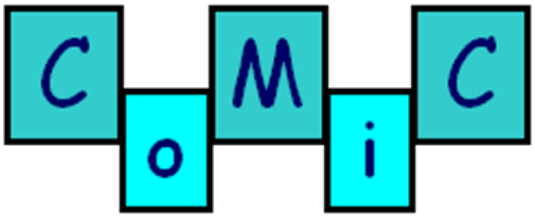
Author: Mary Ellen Foster Date: 27 September 2004

Deliverable 6.5
User evaluation of generated deictic gestures in the T24 demonstrator



Document history

Date	Editor	Explanation	Status
23 August	MEF	Initial	Draft
31 August	MEF	Mike's comments; more stats	Draft
3 September	MEF	More revisions	Public draft
27 September	MEF	Final version	Final

COMIC

Information sheet issued with Deliverable 6.5

Title:	User evaluation of generated deictic gestures in the T24 demonstrator
Abstract:	We describe a study designed to compare user evaluations of the mouse-pointer gestures planned by the the two versions of the T24 fission module. The main result of this study is that subjects were generally able to tell the difference between the gestures planned by the two modules, and largely preferred those generated by the rule-based module to those from the stochastic module. At the end of this report, we describe future plans for the fission module that take into account these findings.
Author(s):	Mary Ellen Foster
Reviewers:	Michael White, Jon Oberlander
Project:	COMIC
Project number:	IST-2001-32311
Date:	27 September 2004
	Public
Key words:	

Distribution list

COMIC partners:	Edinburgh, DFKI, KUN, MPI-N, MPI-T, Sheffield, ViSoft
External COMIC:	PO, reviewers

Contents

> It makes me feel as though version 2 is trying to be more human-like in its actions, though since I'm very aware that it's merely a computer, it feels a bit wrong.
>
> Comment from subject #26

1 Introduction

The fission module is the component in the COMIC demonstrator that plans and produces output based on messages from the dialogue manager. There are two versions of the T24 fission module, which use different techniques to choose the multimodal components of the presentation. The main difference between the modules lies in how they select the deictic "gestures" of the on-screen pointer. Both use the results from an annotated corpus of dialogues to make the decisions: one version uses deterministic rules based on the corpus, while the other makes stochastic decisions using weights derived from the corpus. The details of the corpus and how the data is used in the modules are provided in Section 2.2.

Note that the two versions are called *sloppy* and *strict* in the technical annex and in the descriptions of the implemented modules (Foster, 2004b,c); however, since those terms are confusing given the actual difference between the versions, in this document we will instead use the labels *rule-based* and *stochastic*.

We were interested in determining whether subjects could tell the difference between the output generated by the two versions, and if so, which type of output they preferred. To address these questions, we performed the user study described in the remainder of this document. The design of this study is outlined in Section 3, while the results are described in Section 4.

To summarise the results, the subjects were generally able to tell the difference between the two types of output, and largely preferred the rule-based version to the stochastic one. In Section 5, we propose explanations for this preference and outline the future plans for multimodal planning in COMIC taking into account these results.

2 Planning deictic gestures in the COMIC demonstrator[1]

One of the output channels in the COMIC demonstrator is deictic "gestures" with a simulated mouse pointer at objects on the screen. In the current fission modules, we first prepare the content of the speech, and then select gestures based on that content, as follows. First, we mark all of the noun phrases in the text that have potential on-screen referents, and also indicate features of each noun phrase including whether it is the first reference to that object and whether the reference is deictic. We then use this list of referents to select the gestures. Finally, we filter the gestures to eliminate any overlap.

There are three decisions that must be made as part of the gesture-planning process. First, we must decide whether to include a gesture or not for each potential referent. Second, once we have decided to include a gesture, we must determine the exact type of gesture to use. Finally, we must choose the relevant timing of the speech and the gesture. This section describes the corpus that

[1] This section is based on (Foster, 2004a).

we used to help make these decisions, and how that corpus data was used in the rule-based and stochastic versions.

2.1 The corpus

The corpus is made up of a number of role-playing interactions, with the participants taking the roles of a sales consultant and a client. We then annotated and analysed the deictic gestures and spoken references occurring in the relevant parts of these interactions. This section gives an overview of how these interactions were recorded and annotated; full details are provided in (Foster, 2003).

The subject playing the consultant was given five to seven possibilities for each choice that the client could make in designing a bathroom, and was instructed to help the client to explore this range of possibilities. Each of the design possibilities was presented on an individual sheet of paper. Seven dialogues were recorded, making a total of two and a half hours of video. About 20% of this time contained descriptions and comparisons from the consultant that were similar to those that COMIC can generate.[2] For example, (1) is a comparison produced by one of our subjects, while (2) is a similar comparison generated by COMIC:

(1) This is a very kind of traditional design, just having them lined up down the wall, whereas this, this is kind of ... a bit more audacious perhaps.

(2) This design features terracotta and dark-red in the colour scheme, while this one has a blue and beige colour scheme.

These relevant descriptions and comparisons were annotated as follows. First, we marked the onset and duration of each deictic gesture and spoken reference made by the consultant in the relevant sections. For each gesture, we then indicated whether it referred to an entire image or to a part of that image, and put it into one of four categories: pointing, waving (repeated pointing), circling or tracing the edges of the referent, or physically moving the entire image. In addition, we created links between the spoken and gestural references to the same object, and among the different references to the same objects.

2.2 Using the corpus data in the fission modules

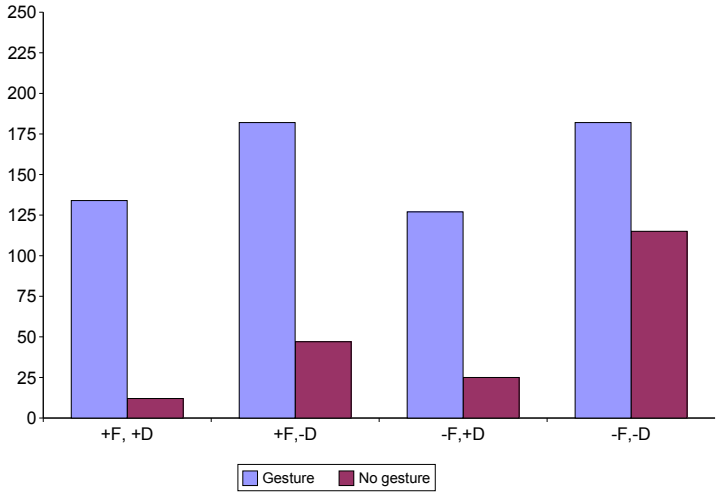
This section describes how we use the data from the corpus to address all of the relevant choices in gesture planning. In the current implementations, these choices are all made independently, as described in Sections 2.2.1–2.2.3. In Section 2.2.4, we describe how we adjust the schedule to eliminate any conflicting gestures.

2.2.1 Deciding whether to include a gesture

The relevant information from the corpus for this decision is the circumstances under which a spoken reference did or did not have an accompanying gesture. We consider two features of a speech reference at this stage: whether it was the first reference to a given entity in the dialogue (F), and whether it was deictic (D). Figure 1 shows the influence of these features on the occurrence

[2] The other 80% consisted mainly of times when the client spoke, and discussion of real-world issues outside the scope of the COMIC demonstrator.

Figure 1: Corpus gesture counts by speech feature



of gestures. Note that both had an influence on the probability of a gesture occurring: 92% of all first, deictic references had a gesture, while only 61% of follow-up, non-deictic references had one. An ANOVA found that all of the differences shown were significant at $p < 0.05$.

The stochastic fission module decides for each noun phrase whether to include an accompanying gesture by making a weighted random choice, using the appropriate probabilities based on the features of that referent. For example, a first referent that is also deictic is accompanied by a gesture with a probability of 0.92. The rule-based module simply makes the majority choice all of the time; since all of the probabilities are above 0.5, this means that a gesture is planned for every potential referent (although overlapping gestures are later filtered as described below).

2.2.2 Choosing the gesture type

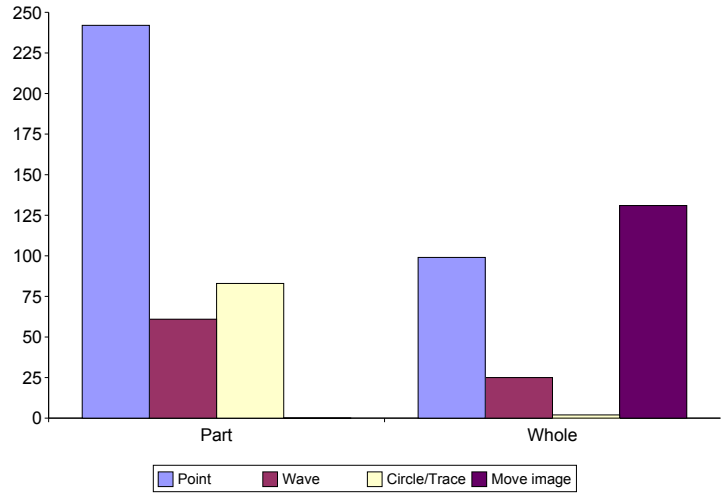
For this choice, the most important factor in the corpus was the object of the gesture—in particular, whether it indicated an entire design or a part of that design (e.g., pointing out the features of specific tiles). Figure 2 shows the distribution of the different gesture types depending on whether the gesture was to a part or a whole. The characteristic gestures to the different object types vary greatly: nearly two-thirds of the gestures to image parts were pointing gestures, while over half of the whole-image gestures involved moving the image. As the animated mouse pointer cannot reproduce all of the observed gesture types, we mapped image moving to circling for this implementation.

As in the previous case, the stochastic module implements this by making a random choice using the weights from the corpus, while the rule-based module uses the majority choice (pointing for a part, circling for a whole).

2.2.3 Choosing an offset

For deictic gestures, the most relevant timing point is the spoken reference to the same object that caused the gesture to be included in the first place. In the recordings, the mean time between

Figure 2: Corpus gesture counts by object type



the onset of a deictic gesture and the onset of a spoken reference to the same object was 0.83 seconds, with a standard deviation of 1.3; that is, a gesture began on average 0.83 seconds before the corresponding speech, although there were some gestures that began a short time after the speech.

The stochastic module reproduces this by choosing the offset from a normal distribution with the same standard deviation and mean, while the rule-based module always starts the gestures exactly synchronised with the speech. In neither version do we currently vary the duration of a gesture.

2.2.4 Resolving conflicts in the schedule

The steps described in the preceding sections produce a preliminary gesture schedule. However, it may be that some gestures in that preliminary schedule overlap with each other. To produce a final, non-conflicting schedule, we perform the following modifications. If two gestures with the same object overlap with each other, we keep only the earlier; if two gestures with different objects overlap, the start time of the second one is modified so that it starts after the first.

3 Experiment design

We performed a user study to compare the success of the gestures planned by the two versions of the module. This study was designed to address the following questions: are users able to notice a difference between the gestures planned by the two versions of the module, and if so, which type of gestures do they prefer?

This section describes the design of this experiment; Section 4 outlines the results of the experiment, while Section 5 analyses these results and describes the plans for using these results in the future development of the fission modules.

3.1 Materials

A total of nineteen outputs were generated in advance, using the full COMIC fission module and realiser. Each output is made up of two parts: a full description of one of the tilesets from the T24 system, and a summary description (style only) of four other randomly-selected tilesets. (3) shows a sample of the textual content of one of the outputs that was used; the subscripts indicate the ID of the tileset being described.

(3) [This design]$_4$ is modern. It uses tiles from Aparici's Carioca series. The colours are white, orange, red, and ochre. There are geometric shapes on the decorative tiles. [This design]$_3$ is in the classic style, while [this one]$_9$ is in the modern style. [This design]$_{11}$ is country, while [this one]$_{21}$ is modern.

Two multimodal scripts were prepared for each description, one with gestures created by each version of the module. The concrete schedules for each description were stored and played back, so that every subject saw exactly the same versions of each. Appendix A shows all of the scripts that were used.

3.2 Subjects

The subjects for this study were the same as those used in the evaluation of the entire COMIC system; this study was done immediately after the whole-system evaluation (TODO cite deliverable here), so the subjects were already familiar with the COMIC system. There were 35 subjects in total for this evaluation: 13 female, and 22 male. All were native speakers of English—mostly (30 of them) of some dialect of British English. The results from two additional subjects were discarded because it was evident from their responses that they had not properly understood the experiment. The initial answers from two other subjects were also discarded, because those subjects did not initially understand.

3.3 Procedure

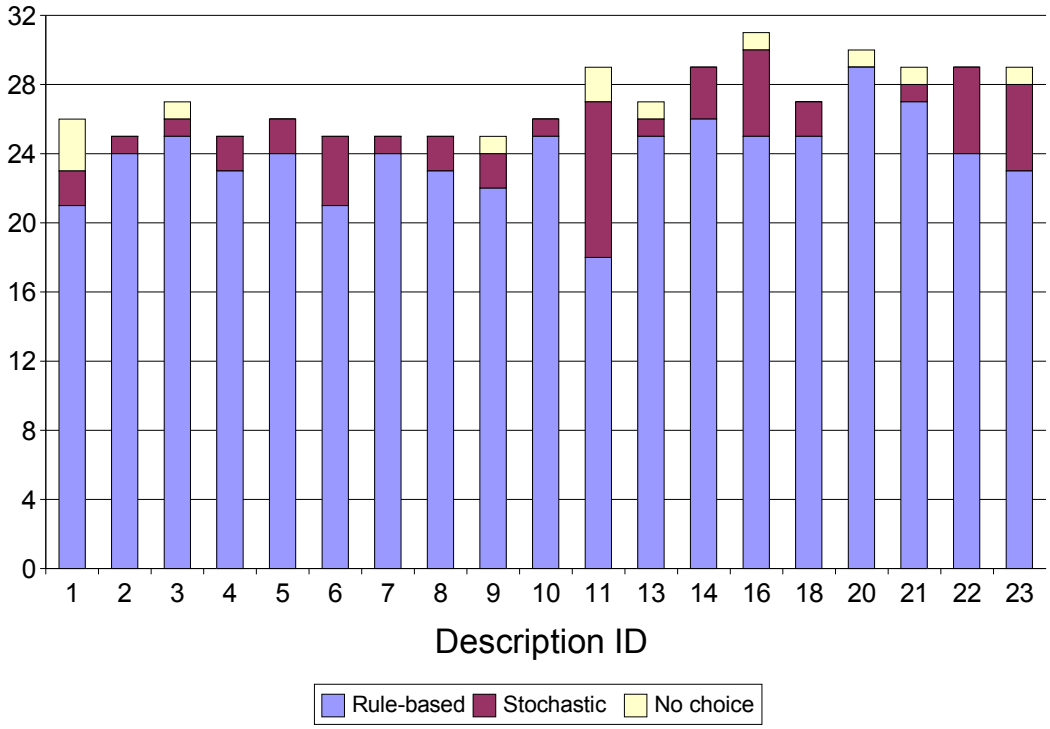
The subjects were shown the two versions of each output and then asked to choose which of the two versions they preferred, giving a reason for each choice if possible. The order of the descriptions was selected randomly for each subject. The order of the two versions of each description was also selected randomly. The instructions and questionnaire that were used are shown in Appendix B.

There were a total of 19 descriptions that could be viewed—one description of each tileset in the T24 COMIC system. Due to time constraints, not all subjects were able to view all of the descriptions. Sixteen subjects were able to view all 19 descriptions; the minimum number of descriptions seen was 8, and the average 14.7. Because not all subjects saw all of the descriptions, there was a slight variation in the number of times each description was viewed. The average number of views of a description was 27, with a maximum of 31 and a minimum of 25.

4 Results

For the most part, the subjects preferred the rule-based gestures to the stochastic gestures. Overall, subjects chose the rule-based gestures 88% of the time and the stochastic gestures about 10% of

Figure 3: Choices grouped by description ID



the time. The remaining 2% of items were those on which subjects failed to make a choice; except when explicitly noted, we will discard those items in the analysis that follows.

Section 4.2 discusses the reasons given by the subjects for their choices, while Section 4.1 breaks down the choices by item, by subject, and by presentation order.

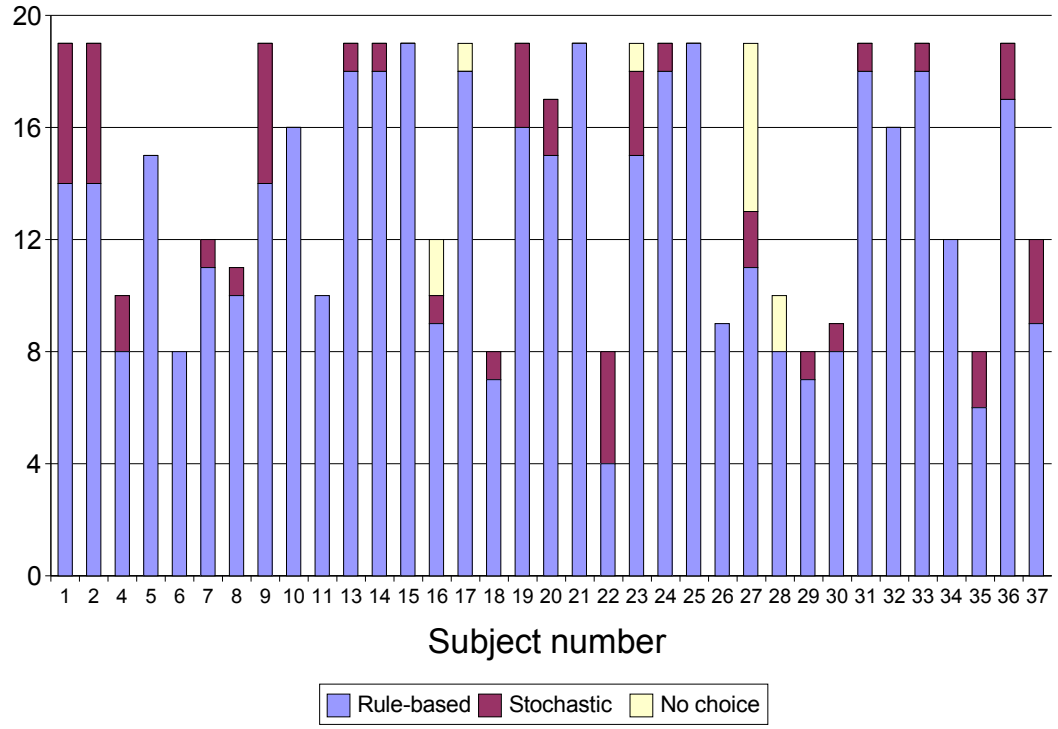
4.1 Breakdown of choices

We can group the choices made by the subjects in three different ways: by the ID of the description, by the subject, or by presentation order.

By description ID Figure 3 shows a graph of the choices made for each description ID. The percentage of subjects choosing the description with rule-based gestures ranged from 100% (description 20; one subject failed to make a choice) to 62% (description 11); see Appendix A for the scripts used for all of the descriptions, and the percentage of subjects choosing the stochastic version of each. Using a binomial test, almost all of the preferences were significant at least at $p < 0.001$—in some cases, p was much smaller. The only exception was description 11: for this item, the preference was still significant, but only at $p < 0.05$.

By subject Figure 4 shows the choices made by each subject; recall that, due to time constraints, not all subjects saw all 19 of the descriptions. 10 subjects chose the rule-based version 100% of the

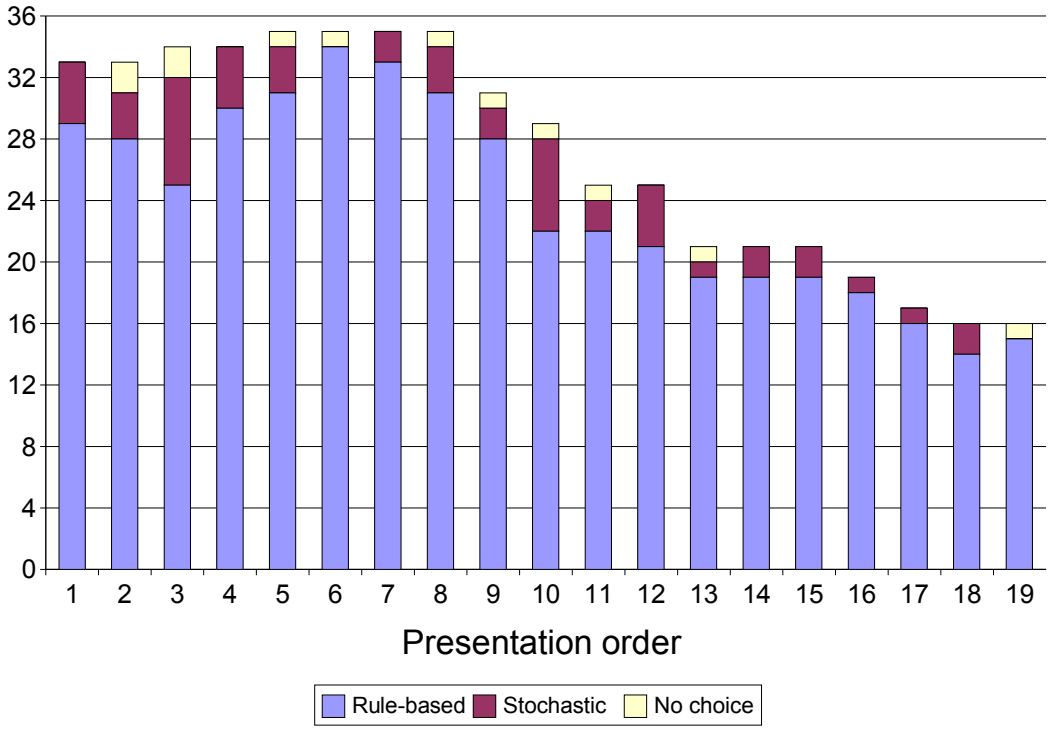
Figure 4: Choices grouped by subject



time; on the other extreme, subject 22 split the choice 50-50 between rule-based and stochastic. Five subjects failed to make a choice on at least one output; subject 27 failed to make a choice on six outputs.

By presentation order Figure 5 shows the choices made, grouped by the order in which the descriptions were presented. Note that the totals for positions 1–4 are less than 35 because the initial answers for two subjects that did not immediately get the point of the experiment were discarded. There is no real effect of presentation order on the choices made.

Figure 5: Choices grouped by presentation order



4.2 Reasons for choices

The reasons subjects gave for making their choices fell into several main categories. Table 1 lists the categories, with examples of each type; Figure 6 shows the number of times each reason was given as a justification for choosing each description type. Note that some comments fell into more than one category.

As can be seen from Figure 6, the distribution of reasons is different for the two choices. The rule-based versions were chosen primarily because they were well-synchronised with the speech timing, pointed at all of the objects on the screen, and had fewer instances of anomalous or jerky output. The biggest class of reasons for choosing the stochastic versions, in contrast, was *other*—most of these reasons were vague explanations such as "more interesting". The proportion of stochastic versions chosen for clarity and naturalness were also higher.

Table 1: Reason categories

Synch Comments that emphasise the synchronisation (or lack thereof) between the speech and the gestures.

- "Pen more in sync with the voice"
- "Timing slightly off on the first one"
- "Less jumpy, more insync with vocals"

Object Comments referring to the objects of the pointing gestures.

- "The pen didn't circle all of the tile groups"
- "It pointed to each tiles when describing it whereas the first version only pointed to some"
- "It 'pointed' things out more for the first design"

Bug Comments pointing out anomalies or lack of smoothness in the gesture output.

- "version 2 looks like it had 'crashed', flashing etc"
- "Second pen in the first one"
- "Pointer didn't match description and jumped about"

Clear Comments referring to how clear or confusing a description was to follow.

- "Clearer and more concise"
- "I was completely lost with version 2!"

Natural Comments pointing out the naturalness of the pointer movements.

- "First one completely natural, second one movement was awkward"
- "v2's up/down motion seemed artificial"

Consist Comments referring to the consistency of the gestures used.

- "v2 didn't circle or point consistently"
- "inconsistency in manner of pointing was annoying"

Pref Comments where subjects expressed a preference for a particular style of gesturing.

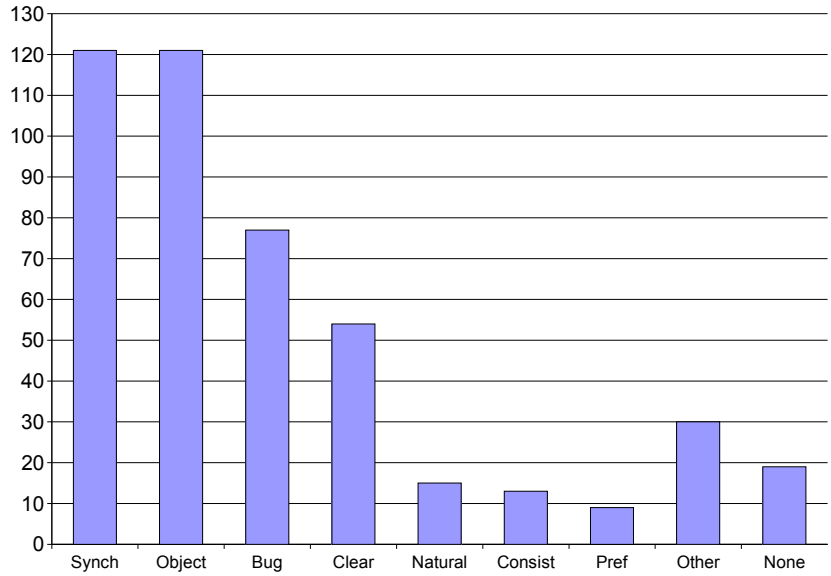
- "The pen moving across the name of the tiles makes it difficult to read."
- "Version 2 moved too much and pointed to too little"

Other Other reasons; mostly referring to factors having nothing to do with the gestures themselves.

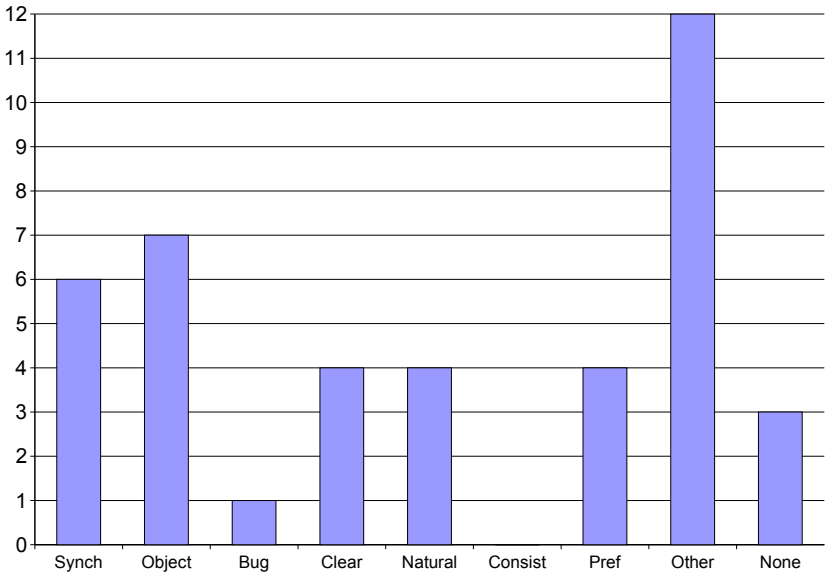
- "was more interesting" *(This was the sample explanation; see Appendix B.)*
- "it seemed to reiterate the first design better"

None No reason given

Figure 6: Reason counts for each description type



(a) Rule-based



(b) Stochastic

Figure 7: Scripts for Description 11

Speech	Rule	Stochastic	
This design is country.	Circle	Point	2572
The tiles are from …	Point	Point	985
… *the Smart collection by Gardenia Orchidea*.	Point	Point	891
The colours are *white and green*.	Point	Circle	-2004
It has *geometric shapes* on the decorative tiles.	Point	Point	1571
These designs are in the classic style, …	Circle	Circle	312
… while *this one* is in the family style.	Circle	Circle	1109
This design is in the classic style.	Circle	Circle	-1487

5 Discussion

The main result presented in the preceding section is that, however the results are grouped, the subjects overwhelmingly preferred the descriptions generated by the rule-based module to those generated by the stochastic module. Their main reasons given for preferring those descriptions were that the rule-based versions were better synchronised with the speech, and that there were always gestures at the thumbnails during the second half of the description.

One factor that could have had an influence on the strength of this preference is that all of the subjects performed this study immediately after an extended interaction with the full COMIC system in a configuration that included the rule-based fission module. This could have led them to prefer the rule-based versions even more strongly than they would otherwise have done. However, that preferences expressed were so strong that the familiarity factor likely only increased its strength; the main message is still that subjects preferred the rule-based versions.

5.1 Description 11

The choices for description 11 were very different than those for the other descriptions. For this one, only 67% of the subjects chose the rule-based version, while the lowest percentage for the rest of the descriptions was 84% and the majority were above 90%. The scripts for Description 11 are reproduced in Figure 7.

Several features contributed to this difference. Firstly, the gesture sequence in the stochastic version is almost the same as in the rule-based version; there are only two differences, where pointing is exchanged with circling. In particular, the stochastic version uses a circling gesture for all of the deictic gestures to the thumbnails.

Secondly, while this description does have some large offsets, in almost all cases, the subjective effect is that the gesture is well coordinated with a different, coreferential noun phrase in the same sentence. For example, the gesture that is planned to coincide with *geometric shapes* in the fourth sentence will appear to be coordinated with *on the decorative tiles* in the output.

5.2 Next steps

The main characteristics of the output that were particularly unpopular with subjects were the following:

- A mixture of positive and negative offsets, even within the same sentence, and very large offsets.
- Gesturing at only some of the thumbnails in the description.
- Mixing pointing with circling in a sequence of gestures.

These problems arose from two main issues in the implementation. First, all of the decisions were made independently, with no contextual information. Secondly, the way the gesture timing was selected used a fairly crude approximation of the corpus data: it may be that the different types of gestures had different characteristic offsets, and that combining them all into a single normal distribution resulted in a standard deviation that was higher than it should have been, thus producing the large offsets.

In the next version of the fission modules, we intend to address these deficiencies in several ways. In cases where it is straightforward to write a rule, we will use one—for example, rules may be used to ensure that all of the thumbnails receive a gesture. However, particularly when we add facial expressions to the repertoire of system output, it is likely that there will be decisions for which rules will not be as easy to write.

In these cases, we intend to employ instance-based techniques to help plan the presentations, as follows. We will use the *n*-gram language models of the text realiser (White, 2004a,b) to choose among *multimodal n*-grams, where information about the gestures and facial expressions accompanying the words is incorporated into the realiser's search for high-scoring surface forms. This should result in output without the inconsistencies that the stochastic module showed here, but that is still more varied natural than that of the rule-based module. We will perform a similar evaluation of the updated modules to determine whether an instance-based module produces output that is preferred over that of a rule-based module.

6 References

FOSTER M E (2003). *Description of the "Wizard of Oz" recordings*. Public deliverable 6.4, COMIC project.

FOSTER M E (2004a). Corpus-based planning of deictic gestures in COMIC. In: *Proceedings of INLG-04 student session*.

FOSTER M E (2004b). *Sloppy fission module (T22)*. Internal deliverable 6.2, COMIC project.

FOSTER M E (2004c). *Strict fission module (T22)*. Internal deliverable 6.3, COMIC project.

WHITE M (2004a). Efficient realization of coordinate structures in Combinatory Categorial Grammar. *Research on Language and Computation* To appear.

WHITE M (2004b). Reining in CCG chart realization. In: *Proceedings of INLG 2004*.

A Outputs

The following are the scripts that were used in this experiment. The percentage in each header indicates the proportion of subjects who chose the rule-based version of that description. The italicised words in the text indicate the noun phrases with which gestures were coordinated. The *Rule* column shows the gesture that was included in the rule-based script for that NP (if any)—all offsets were zero in those versions. The *Stochastic* column shows the gesture and offset that were included in the stochastic script.

Videos of both versions of the output for descriptions number 18 and 22 are available on the web at http://www.hcrc.ed.ac.uk/comic/demos/t24-gestures/.

Description 1 (91%)

Speech	Rule	Stochastic	
This design is country	Circle	Circle	-1008
It uses tiles from *Villeroy and Boch's Altamira series*.	Point	Point	1033
There are *floral motifs* on the decorative tiles.	Point	Circle	-1201
The tiles have a sandstone, green and blue colour scheme.	Point	Circle	62
These designs are classic, ...	Circle	Circle	1297
... while *these two* are in the modern style.	Circle	Circle	1779

Description 2 (96%)

Speech	Rule	Stochastic	
This design is classic.	Circle	—	—
The tiles are from ...	Point	Point	-3532
... *the Armonie collection by Iris Ceramica*.	Point	—	—
The deep glaze effect is the distinguishing mark of this collection.	—	—	—
It features *white, yellow and rose* in the colour scheme.	Point	Point	1319
There are *floral motifs* on the decorative tiles.	Point	Wave	850
These designs are in the classic style.	Circle	Circle	-1290

Description 3 (96%)

Speech	Rule	Stochastic	
This design is classic.	Circle	Point	-506
It's based on *the Abbazie collection by NovaBell*.	Point	—	—
It has *geometric shapes* on the decorative tiles.	Point	—	—
The colours are *beige and ochre*.	Point	Point	2471
This design is ...	Circle	—	—
... *in the country style*, ...	—	Point	1868
... while *this one* is in the classic style.	Circle	Point	1338
This design is in the country style, ...	Circle	Circle	-1151
... while *this one* is in the classic style.	Circle	Circle	2112

Description 4 (92%)

Speech	Rule	Stochastic	
This design is modern.	Circle	Circle	1000
It uses tiles from *Aparici's Carioca series*.	Point	Wave	1570
The colours are *white, orange, red and ochre*.	Point	Point	-843
There are *geometric shapes* on the decorative tiles.	Point	Point	-485
This design is in the classic style, …	Circle	Circle	1017
… while *this one* is in the modern style.	Circle	Circle	2599
This design is country, …	Circle	—	—
… while *this one* is modern.	Circle	Circle	2056

Description 5 (92%)

Speech	Rule	Stochastic	
This design is in the modern style.	Circle	Point	3731
It's based on the *Helenus collection by Sphinx Tiles*.	Point	Point	-3103
The tiles have a white and black colour scheme.	Point	Point	52
There are *geometric shapes* on the decorative tiles.	Point	Point	692
This design is family, …	Circle	—	—
… while *this one* is in the modern style.	Circle	Wave	162
This design is classic, …	Circle	Circle	425
… while *this one* is in the modern style.	Circle	Wave	1439

Description 6 (84%)

Speech	Rule	Stochastic	
This design is in the classic style.	Circle	Circle	650
It uses tiles from *Porcelanosa's I Marmi series*.	Point	Wave	2105
There are *geometric shapes* on the decorative tiles.	Point	Wave	1071
It features *white and dark green* in the colour scheme.	Point	Circle	1409
These designs are in the classic style.	Circle	Circle	147
This design is in the modern style.	Circle	Point	3336

Description 7 (96%)

Speech	Rule	Stochastic	
This design is classic.	Circle	Wave	-645
It's based on *the Jazz collection by Porcelaingres*.	Point	Point	1018
The white, red and yellow colour scheme emphasises the clear, straight lined character of your bathroom in a stylish way.	Point	Wave	-645
It has *geometric shapes* on the decorative tiles.	Point	Circle	1246
These designs are classic, …	Circle	Point	-807
… while *this one* is country.	Circle	—	—
This design is classic.	Circle	—	—

Description 8 (92%)

Speech	Rule	Stochastic	
This design is classic.	Circle	Circle	2023
The tiles draw from *Levante, by Cerim Ceramiche*.	Point	—	—
It has *abstract shapes* on the decorative tiles.	Point	Point	1726
The tiles have a red, pink and beige colour scheme.	Point	Point	938
This design is country, …	Circle	—	—
… while *this one* is in the modern style.	Circle	Wave	2226
These designs are classic.	Circle	Circle	312

Description 9 (92%)

Speech	Rule	Stochastic	
This design is in the modern style.	Circle	Point	2534
It uses tiles from *Apavisa's Palace series*.	Point	Circle	-598
The tiles have a beige, cream and salmon colour scheme.	Point	Point	633
It has *geometric shapes* on the decorative tiles.	Point	Circle	2094
This design is in the classic style, …	Circle	Circle	2041
… while *this one* is country.	Circle	Circle	1922
This design is …	Circle	—	—
… *in the family style*, …	—	Wave	862
… while *this one* is classic.	Circle	—	—

Description 10 (96%)

Speech	Rule	Stochastic	
This design is in the classic style.	Circle	Point	1220
The tiles are from …	Point	Point	-483
… *the Tirrenia collection by Imola Ceramica*.	Point	Wave	-437
There are *floral motifs and fruit* …	Point	—	—
… *on the decorative tiles*.	—	Circle	1831
The tiles have a white and blue colour scheme.	Point	Point	1657
This design is in the classic style, …	Circle	Circle	1177
… while *this one* is modern.	Circle	—	—
This design is in the classic style, …	Circle	Circle	1953
… while *this one* is modern.	Circle	Point	485

Description 11 (67%)

Speech	Rule	Stochastic	
This design is country.	Circle	Point	2572
The tiles are from …	Point	Point	985
… *the Smart collection by Gardenia Orchidea*.	Point	Point	891
The colours are *white and green*.	Point	Circle	-2004
It has *geometric shapes* on the decorative tiles.	Point	Point	1571
These designs are in the classic style, …	Circle	Circle	312
… while *this one* is in the family style.	Circle	Circle	1109
This design is in the classic style.	Circle	Circle	-1487

Description 13 (96%)

Speech	Rule	Stochastic	
This design is in the classic style.	Circle	Point	758
It's based on *the Alt Mettlach collection by Villeroy and Boch*.	Point	Point	-66
This is one of our most popular collections.	—	—	—
It has *floral motifs and geometric shapes* on the decorative tiles.	Point	Wave	3780
The colours are *off white, terracotta and black*.	Point	Circle	721
This design is classic, …	Circle	Point	796
… while *this one* is in the modern style.	Circle	Circle	2833
These designs are classic.	Circle	Point	1390

Description 14 (90%)

Speech	Rule	Stochastic	
This design is in the modern style.	Circle	Circle	2020
Its tiles are from …	Point	Point	-667
… *the Century Esprit collection by Villeroy and Boch*.	Point	—	—
There are *floral motifs and abstract shapes* on the decorative tiles.	Point	Point	771
The floral motifs are combined with abstract shapes in a playful way.	—	—	—
The colours are *blue and beige*, helping to set the mood.	Point	Point	249
This design is in the classic style, …	Circle	Circle	269
… while *this one* is modern.	Circle	Point	2146
This design is classic, …	Circle	Point	997
… while *this one* is in the modern style.	Circle	Circle	253

Description 16 (83%)

Speech	Rule	Stochastic	
This design is classic.	Circle	Circle	572
Its tiles are from …	Point	Point	1105
… *the Armonie collection by Iris Ceramica*.	Point	Circle	972
The deep glaze effect is the distinguishing mark of this collection.	—	—	—
The colours are *white and rose*.	Point	—	—
There are *geometric shapes* on the decorative tiles.	Point	Point	1759
This design is in the modern style, …	Circle	Point	446
… while *these two* are in the classic style.	Circle	Circle	1747
This design is in the family style.	Circle	Wave	1707

Description 18 (93%)

Speech	Rule	Stochastic	
This design is in the family style.	Circle	Circle	543
The tiles are from …	Point	Point	1714
… *the Hippo collection by Engers Eurodesign*.	Point	Point	2376
There's *artwork* on the decorative tiles.	Point	Circle	4310
The hippos are always a hit with the kids.	—	—	—
The tiles have a blue and green colour scheme.	Point	Point	1145
This design is in the country style, …	Circle	Point	437
… while *this one* is modern.	Circle	Circle	1318
This design is in the classic style, …	Circle	Circle	3126
… while *this one* is in the modern style.	Circle	Point	1658

Description 20 (100%)

Speech	Rule	Stochastic	
This design is in the family style.	Circle	Circle	662
The tiles are from …	Point	Point	2435
… *the Lollipop collection by Agrob Buchtal*.	Point	Point	2278
The tiles have a blue and green colour scheme.	Point	Point	-638
It has *floral motifs and artwork* on the decorative tiles.	Point	Point	586
The lollipops are always a treat for the kids.	—	—	—
This design is in the country style, …	Circle	Circle	710
… while *this one* is modern.	Circle	—	—
These designs are in the classic style.	Circle	Circle	-301

Description 21 (97%)

Speech	Rule	Stochastic	
This design is modern.	Circle	Point	107
Its tiles are from …	Point	Circle	-1657
the HundertWasser collection by Steuler.	Point	Point	2734
It has *artwork* on the decorative tiles.	Point	Point	2664
It features *white, black and blue* in the colour scheme.	Point	Point	843
This design is modern, …	Circle	Point	1388
… while *this one* is in the family style.	Circle	Circle	316
These designs are classic.	Circle	Circle	2304

Description 22 (83%)

Speech	Rule	Stochastic	
This design is in the modern style.	Circle	Circle	1192
It uses tiles from *Villeroy and Boch's Century series*.	Point	—	—
The tiles and decorative motifs come from a culture which has taken ceramic design to a fine art.	—	—	—
It has *floral motifs and geometric shapes* …	Point	—	—
… *on the decorative tiles*.	—	Circle	1900
The colours are *blue, beige, terracotta and white*.	Point	Circle	2827
These designs are in the classic style, …	Circle	Circle	-43
… while *this one* is family.	Circle	Point	1744
This design is in the modern style.	Circle	Circle	-526

Description 23 (82%)

Speech	Rule	Stochastic	
This design is in the classic style.	Circle	Point	2632
It uses tiles from *Bisazza's Opus Romano series*.	Point	Circle	1082
The colours are *black, white and beige*.	Point	Point	1405
There are *mosaics* on the decorative tiles.	Point	Circle	1589
Mosaic tiles are very hard wearing.	—	—	—
This design is in the family style, …	Circle	Circle	-603
… while *these two* are in the modern style.	Circle	Circle	3010
This design is in the modern style.	Circle	Circle	1572

B Instructions and questionnaire

The following pages show the instructions and questionnaire that were given to the subjects in this evaluation. The instructions start with “For the last part of this experiment” because all of the subjects had previously used the full system. Due to time constraints, not all subjects evaluated all nineteen descriptions.

For the last part of the experiment, we are going to show you another nineteen descriptions generated by the system. You will see two versions of each description, which use different ways of pointing to things on the screen. For each pair, we would like you to indicate which of the versions you liked better, and describe any reasons for your choice.

You will see the descriptions one pair at a time. After each pair, you should check off the box corresponding to the version that you liked better, and describe in the space below why you chose that version. You must choose one of the versions; please try to give a reason for every choice if possible. When you have made your choice, let us know and we will play the next set of descriptions.

For example, if you thought that the first version of description 5 was better because it was more interesting, you would answer like this:

Description 5	**Version 1**	**Version 2**
Which version was better?	✔	❍

Why did you choose this version?

It was more interesting.

Please choose one version for each description, and describe why if possible.

Description 1	**Version 1**	**Version 2**
Which version was better?	❍	❍

Why did you choose this version?

Description 2	**Version 1**	**Version 2**
Which version was better?	❍	❍

Why did you choose this version?

Description 3	**Version 1**	**Version 2**
Which version was better?	❍	❍

Why did you choose this version?

Description 4	**Version 1**	**Version 2**
Which version was better?	❍	❍

Why did you choose this version?

Please choose one version for each description, and describe why if possible.

Description 5	**Version 1**	**Version 2**
Which version was better?	❍	❍

Why did you choose this version?

Description 6	**Version 1**	**Version 2**
Which version was better?	❍	❍

Why did you choose this version?

Description 7	**Version 1**	**Version 2**
Which version was better?	❍	❍

Why did you choose this version?

Description 8	**Version 1**	**Version 2**
Which version was better?	❍	❍

Why did you choose this version?

Please choose one version for each description, and describe why if possible.

Description 9	**Version 1**	**Version 2**
Which version was better?	❍	❍

Why did you choose this version?

Description 10	**Version 1**	**Version 2**
Which version was better?	❍	❍

Why did you choose this version?

Description 11	**Version 1**	**Version 2**
Which version was better?	❍	❍

Why did you choose this version?

Description 12	**Version 1**	**Version 2**
Which version was better?	❍	❍

Why did you choose this version?

Please choose one version for each description, and describe why if possible.

Description 13	**Version 1**	**Version 2**
Which version was better?	❍	❍

Why did you choose this version?

Description 14	**Version 1**	**Version 2**
Which version was better?	❍	❍

Why did you choose this version?

Description 15	**Version 1**	**Version 2**
Which version was better?	❍	❍

Why did you choose this version?

Description 16	**Version 1**	**Version 2**
Which version was better?	❍	❍

Why did you choose this version?

Please choose one version for each description, and describe why if possible.

Description 17	**Version 1**	**Version 2**
Which version was better?	❍	❍

Why did you choose this version?

Description 18	**Version 1**	**Version 2**
Which version was better?	❍	❍

Why did you choose this version?

Description 19	**Version 1**	**Version 2**
Which version was better?	❍	❍

Why did you choose this version?
