Corpus-Based Generation of Head and Eyebrow Motion for an Embodied Conversational Agent *

Mary Ellen Foster (foster@in.tum.de)
Informatik VI: Robotics and Embedded Systems
Technische Universität München

Jon Oberlander (jon@inf.ed.ac.uk)
Institute for Communicating and Collaborative Systems
School of Informatics, University of Edinburgh

Abstract. Humans are known to use a wide range of non-verbal behaviour while speaking. Generating naturalistic embodied speech for an artificial agent is therefore an application where techniques that draw directly on recorded human motions can be helpful. We present a system that uses corpus-based selection strategies to specify the head and eyebrow motion of an animated talking head. We first describe how a domain-specific corpus of facial displays was recorded and annotated, and outline the regularities that were found in the data. We then present two different methods of selecting motions for the talking head based on the corpus data: one that chooses the majority option in all cases, and one that makes a weighted choice among all of the options. We compare these methods to each other in two ways: through cross-validation against the corpus, and by asking human judges to rate the output. The results of the two evaluation studies differ: the cross-validation study favoured the majority strategy, while the human judges preferred schedules generated using weighted choice. The judges in the second study also showed a preference for the original corpus data over the output of either of the generation strategies.

Keywords: data-driven generation; embodied conversational agents; evaluation of generated output; multimodal corpora

1. Introduction

It has long been documented that the verbal and non-verbal components of embodied speech are tightly linked. For example, Ekman (1979) noted that eyebrow raises "appear to coincide with primary voice stress, or more simply with a word that is spoken more loudly." Similarly, Graf et al. (2002) found that in their corpus of facial recordings, "rises of eyebrows are often placed at prosodic events, sometimes with head nods, at other times without." However, while correlations have been found between facial displays and prosodic events, this is not a strict rule: in normal embodied speech, many pitch accents and other prosodic

* This work was supported by the EU projects COMIC (IST-2001-32311) and JAST (FP6-003747-IP). An initial version of this study was published as Foster and Oberlander (2006).

events are unaccompanied by facial displays, while other facial displays occur with no obviously related prosodic event. Other factors including information structure, syntactic structure, and affective and pragmatic context can also influence a speaker's non-verbal behaviour.

Since there are so many factors that can influence the non-verbal behaviours that accompany speech, specifying appropriate multimodal behaviour for an artificial embodied agent is a complex task—and one where models derived from recorded human data can be helpful. In this project, we select the head and eyebrow motions of a synthetic talking head in a multimodal dialogue system based on the recorded and annotated behaviour of a speaker reading a script of similar sentences. We implement two different selection techniques, majority choice and weighted choice, and compare them using two methods: by computing a range of automated corpus similarity measures, and by gathering the opinions of human judges.

By building and evaluating models of multimodal human behaviour based on manual annotation of a video corpus of a human speaker, the work described in this paper contributes to the growing body of knowledge about multimodal behaviour. The corpus is used to generate the behaviours of an embodied conversational agent, and also to evaluate those behaviours. The conclusions concerning the relative utility of cross-validation and human evaluation for generation contribute to the emerging consensus that the latter is absolutely essential.

2. Background

This study builds on work in three areas: generating non-verbal behaviour for embodied conversational agents, building and using multimodal corpora, and using corpora in generation systems. In this section, we summarise the main techniques and issues in each of these areas.

2.1. Embodied conversational agents

An Embodied Conversational Agent (ECA) is a computer interface that is represented as a human body, and that uses its face and body in a human-like way in conversation with the user (Cassell et al., 2000). The main benefit of an ECA as a user-interface device is that it allows users to interact with a computer in the most natural possible setting: face-to-face conversation. However, to take full advantage of this benefit, the conversational agent must produce high-quality output, both verbal and non-verbal. Non-verbal behaviour has two main aspects: motions such as beat gestures and emphatic facial displays that correspond

directly to the structure of the speech, and other behaviours such as emotional facial expressions that are related to the pragmatic context.

An ECA system generally uses the recorded behaviour of humans in conversational situations to choose the motions of the agent. There are two main implementation strategies. In some cases, the recorded behaviours are analysed by hand and rules are created to make the selection; in others, models based directly on the recorded data are used the decision process. The performative facial displays for the Greta agent (de Carolis et al., 2002), for example, were selected using the former technique: rules to map from emotional states to facial displays were derived from the literature on facial expressions of emotion. Similarly, Cassell et al. (2001a) selected gestures and facial expressions for the REA agent using heuristics derived from studies of typical North American non-verbal displays. An implementation of this sort tends to produce averaged behaviours from a range of speakers, but does not include specific personality and stylistic effects, and tends to draw from a small range of alternative behaviours.

In contrast, the non-verbal behaviour of other ECAs is selected using models built directly from the data; such systems are able to produce more naturalistic output than a rule-based system, and can also easily model a single individual. Stone et al. (2004), for example, used motion capture to record an actor performing scripted output in the domain of a computer game. They segmented the recordings into coherent phrases and annotated them with the relevant semantic and pragmatic information, and then combined the segments at run-time to produce performance specifications to be played back on an embodied agent. Similarly, Mana and Pianesi (2006) captured the facial motions of an actor speaking nonsense words in a range of emotional contexts, modelled the behaviour using a hidden Markov model, and then used the model to specify MPEG-4 animation commands for a talking head.

Both of the above systems used corpora of human non-verbal behaviour built using automated motion capture; this requires specialised hardware and software. An alternative strategy is to use manual annotation to create the corpus. Annotating a video corpus can be less technically demanding than capturing and directly re-using real motions, especially when the corpus and the number of features under consideration are small. For example, Cassell et al. (2001b) used this technique to choose posture shifts for the REA agent based on the annotated behaviours of speakers describing a house and giving directions. More recently, Kipp (2004) used a similar technique to generate agent gestures based on annotated recordings of skilled public speakers.

2.2. Multimodal corpora

A multimodal corpus is an annotated collection of coordinated content on communication channels such as speech, gaze, hand gesture, and body language, and is generally based on recorded human behaviour. At the moment, multimodal corpora are primarily employed in descriptive tasks such as analysis and summarisation (Martin et al., 2006); however, they have also been used as resources for making decisions when generating output. In particular, the data in such a corpus can be useful for selecting the behaviour of an embodied agent, as in several applications described in the preceding section.

If a multimodal corpus is to be used for generation, the annotated data must correspond to the inputs and outputs that will be used in the system. This imposes additional requirements on the corpus that do not exist if the primary purpose is analysis. First, the pragmatic context under which each item of the corpus was created must be known; that is, the corpus must include all contextual information that the generator might use to choose among alternatives in a given situation. Also, the content on the different channels must be linked to each other so that the generator can produce properly coordinated output. In some cases, the common strategy of annotating each modality on a separate channel and leaving the links implicit in the temporal information is adequate; however, the temporal relationship among communicative modalities can be complex (cf. McNeill (2000)), so explicit links may be necessary. Finally, the annotated content in the corpus must be described at a level that is appropriate for specifying the output of the target generation system. This level can vary widely: for example, among the data-driven ECA systems mentioned in the preceding section, Kipp (2004) described non-verbal behaviour using a gesture grammar, Mana and Pianesi (2006) used Facial Animation Parameters (FAPs), while Stone et al. (2004) used the captured motions directly.

The necessary correspondence between a multimodal corpus and a generation system can be achieved in several ways. When recording the data, one possibility is to create situations in which the necessary pragmatic context is known in advance so that it does not need to be annotated; this was done, for example, by Stone et al. (2004) and Cassell et al. (2001b). It is also possible to annotate existing recordings to add the contextual information, as was done by Kipp (2004). To obtain compatible input and output specifications and cross-modal links, in most cases the generation system and the annotation scheme are defined in parallel. It is also possible to design a generation system to use the representations found in an existing corpus, but this is not a common strategy.

2.3. Corpora in generation

The increasing availability of large textual corpora has led to increased use of data-driven techniques in many areas of language processing. Researchers in Natural-Language Generation (NLG) have now also begun to make use of such techniques (cf. Belz and Varges (2005)). Modern data-driven NLG systems make use of textual corpora in two ways: on the one hand, corpus data can act as a resource for decision-making at all levels of the generation process, from content determination to lexical choice; on the other hand, the data can also be used to help evaluate the output.

One of the first generation systems to exploit corpus data directly in its decision-making process was Nitrogen (Langkilde and Knight, 1998). Nitrogen works in two stages: first, it maps its semantic inputs into word lattices, and then it uses n-grams derived from text corpora to search the lattice to find the best-scoring realisations. The successor system HALogen (Langkilde-Geary, 2002) adds a fuller treatment of syntax and makes other modifications to permit broader coverage and finer control over the output. Among more recent systems, the OpenCCG surface realiser (White, 2006) uses a chart-based realisation algorithm that ranks edges using n-gram precision scores based on a corpus of target outputs, while many of the ECA systems described in Section 2.1 use data-driven techniques to select non-verbal behaviours.

Corpora have also been used as resources for evaluating the generated output. Although the predictions of metrics based on corpus similarity do not always correspond with the preferences of users (cf. Belz and Reiter (2006)), they do provide a fast and often useful form of evaluation. Bangalore et al. (2000) evaluated the Fergus realisation module using a number of metrics that compared the output of their system directly to the corpus that it was trained on, using either the surface strings or the syntactic trees. They found that the metrics corresponded well with human judgements of word-ordering quality. The current shared-task evaluation campaign for NLG (Belz et al., 2007) includes corpus-based evaluation of generated referring expressions.

3. Building a corpus of non-verbal behaviour

As in many of the systems described in Section 2.1, the goal of the current implementation is to generate naturalistic behaviour for an embodied agent using a model drawn from a corpus of recorded human behaviour. In this section, we describe how the corpus was constructed, recorded, and annotated, and also discuss how it responds to the requirements for a generation corpus.

The implementation is based on the output-generation components (Foster et al., 2005) of the COMIC multimodal dialogue system, which adds a multimodal talking-head interface to a CAD-style system for redesigning bathrooms. We concentrate on the turns where the system describes and compares options for tiling the room, as those are the turns with the most interesting and varied content. An example sentence from this phase of the system is the following description, tailored to the current user's likes and dislikes, of two features of a set of tiles:

(1) Although it's in the family style, the tiles are by Alessi Tiles.

3.1. Recording

The script for the recording consisted of 444 sentences created by the full COMIC output planner, which uses the OpenCCG surface realiser (White, 2006) to create texts including prosodic specifications for the speech synthesiser and incorporates information from the dialogue history and a model of the user's likes and dislikes. Every node in the OpenCCG derivation tree for each sentence in the script was initially labelled with all of the available syntactic and pragmatic information from the output planner, including the following features:

- The user-preference evaluation of the object being described (positive or negative);
- Whether the fact being presented was previously mentioned in the discourse (*as I said before, ...*) or is new information;
- Whether the fact is explicitly compared or contrasted with a feature of the previous tile design (*once again ... but here ...*);
- Whether the node is in the first clause of a two-clause sentence, in the second clause, or is an only clause;[1]
- The surface string associated with the node;
- The surface string, with words replaced by semantic classes or stems drawn from the grammar (e.g., *this design is classic* becomes *this [mental-obj] be [style]*); and
- Any pitch accents specified by the text planner.

Figure 1 illustrates the labelled OpenCCG derivation tree for a sample sentence, where the indentation reflects the derivation structure.

[1] No sentence in the script had more than two clauses.

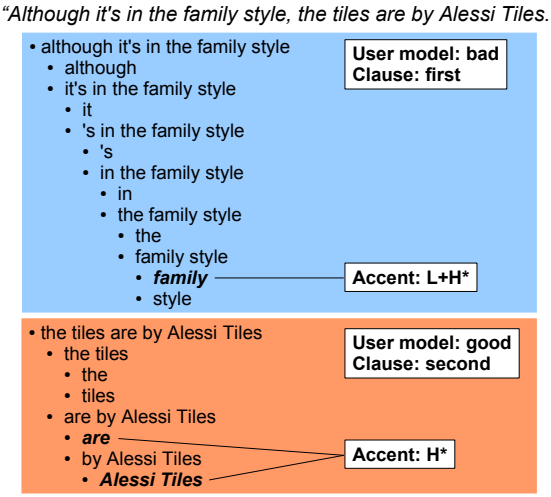

Figure 1. Labelled OpenCCG derivation tree

Every node in the first half of this sentence is associated with a negative user-preference evaluation and is in the first clause of the sentence, while every node in the second half is linked to a positive evaluation and is in the second clause. The figure also shows the pitch accents selected by the output planner according to Steedman's (2000) theory of information structure and intonation.

We recorded a single amateur actor reading all sentences, which were presented one at a time; the presentation included both the textual content (with accented words indicated) and the intended pragmatic context. Each sentence was displayed in a large font on a laptop computer directly in front of the speaker, with the camera positioned directly above the laptop to ensure that the speaker was looking towards the camera at all times. The speaker was instructed to read each sentence out loud as expressively as possible into the camera.

3.2. Annotation

Once the video was recorded, it was split into individual clips corresponding to each sentence using Anvil (Kipp, 2004). We then annotated the speaker's facial displays in each clip, considering five types of motion: eyebrow raising or lowering; eye narrowing; head nodding (up or down), head leaning (left or right), and head turning (left or right). This set of displays was chosen based on the emphatic facial displays documented in the literature, the capabilities of the target talking head, and the actual behaviour of the speaker during the recording session. Figure 2 shows some typical displays.

Figure 2. Typical facial displays

Each display was attached to the span of words that it coincided with temporally. If a single node in the derivation tree exactly covered all of the words spanned by a display, then the annotation was placed on that node; if the words did not coincide with a single node, it was attached to the set of nodes that did cover the necessary words. For example, in the derivation shown in Figure 1, the sequence *the family style* is associated with a single node, so a motion that started and stopped at the same time as that sequence would be attached to the single node. On the other hand, if there were a motion on *the tiles are*, it would be attached to both the *the tiles* node and the *are* node. Any number of displays could be attached to each node.

The annotation tool allowed the coder to play back a recorded sentence at full speed or slowed down, and to associate any combination of displays with any node or set of nodes in the OpenCCG derivation tree of the sentence. The tool also allowed a proposed annotation sequence to be played back on a synthetic talking head to verify that it was

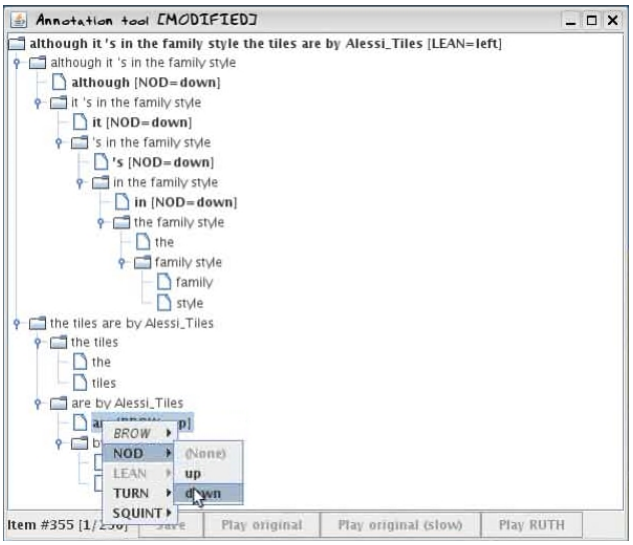

Figure 3. Annotation tool

faithful to the actual motions. Figure 3 shows a screenshot of the annotation tool in use on the sentence from Figure 1. In the screenshot, a left turn has been attached to the entire sentence (i.e., the root node), while a series of nods is associated with single leaf nodes in the first half of the sentence. The coder has already attached a brow raise to the word *are* in the second half and is in the process of adding a downward nod to the same word.

The output of the annotation tool is an XML document including the original labelled OpenCCG derivation tree of each sentence, with each node additionally labelled with a (possibly empty) set of facial displays. Figure 4 shows an excerpt from the annotated version of the sentence from Figure 1. This document includes the full set of features from the original tree. Every node specifies the string generated by the subtree that it spans, both in its surface form (`sf`) and with semantic-class and stem replacement (`sc`). The nodes also have contextual features, indicated by italics in the figure: every node in the second subtree has `um="g"` and `fs="n"` (i.e., a positive evaluation in the second clause), while the accented *are* also has `ac="H*"`. This output tree also includes the facial displays added by the coder in Figure 3, highlighted in the figure by underlining: a left lean (`lean="left"`) attached to the root node and a downward nod (`nod="down"`) accompanied by a brow raise (`brow="up"`) on *are* near the end.

```
<node sf="although it 's in the family style the tiles are by Alessi_Tiles"
    sc="although [pro3n] be in the [style] [abs] the [phys] be by [manuf]"
    lean="left">
  <!-- ... although it is in the family style ... -->
  <node sf="the tiles are by Alessi_Tiles" um="g" fs="n"
      sc="the [phys-obj] be by [manuf]">
    <node sf="the tiles" um="g" fs="n" sc="the [phys-obj]">
      <node sf="the" um="g" fs="n"/>
      <node sf="tiles" sc="[phys-obj]" stem="tile" um="g" fs="n"/>
    </node>
    <node sf="are by Alessi_Tiles" um="g" fs="n" sc="be by [manuf]">
      <node sf="are" stem="be" ac="H*" um="g" fs="n" brow="up" nod="down"/>
      <node sf="by Alessi_Tiles" um="g" fs="n" sc="by [manuf]">
        <node sf="by" um="g" fs="n"/>
        <node sf="Alessi_Tiles" sc="[manuf]" ac="H*" um="g" fs="n"/>
      </node>
    </node>
  </node>
</node>
```

Figure 4. Excerpt from annotated corpus

3.3. Reliability of annotation

Several measures were taken to ensure that the annotation process was reliable. As the first step, two independent coders each separately processed the same set of 20 sentences, using a draft of the annotation scheme. The coders discussed the differences in the outputs and agreed on a final scheme, which one of those coders then used to process the entire set of 444 sentences. As a further test of reliability, an additional coder was trained on the annotation scheme and processed 286 sentences (approximately 65% of the corpus).

To assess the degree of agreement between these two coders, we used a version of the β agreement coefficient proposed by Artstein and Poesio (2005). β is designed as a coefficient that is weighted, that applies to multiple coders, and that uses a separate probability distribution for each coder. Weighted coefficients like β permit degrees of agreement to be measured, so that partial agreement is penalised less severely than total disagreement. Like other weighted coefficients, β is based on the ratio between the observed and expected disagreement on the corpus.

To use this coefficient, we must define a measure that computes the distance between two proposed annotations. We use a measure similar to that proposed by Passonneau (2004) for measuring agreement on set-valued annotations. The full details of the computation are included in Foster (2007); here, we give an informal description.

For each display proposed by each coder on a sentence S, we search for a corresponding display proposed by the other coder—one with the same value (e.g., a brow raise) and covering a similar span of nodes. If both proposals cover the same nodes, that indicates no disagreement (0); if one display covers a strict subset of the nodes covered by the

other, that indicates minor disagreement ($\frac{1}{3}$); if the nodes covered by the two proposals overlap, that is a more major disagreement ($\frac{2}{3}$); and if no corresponding display can be found, there is total disagreement (1). The total observed disagreement $D_o(S)$ is the sum of the disagreement level for each display proposed by each coder on sentence S.

The expected disagreement $D_e(S)$ on a sentence S is based on the length of the sentence. We first use the corpus counts to compute the probability of each coder assigning each possible facial display to word spans of all possible lengths. We then use these probabilities to estimate the likelihood of the two coders assigning identical, super/subset, overlapping, or disjoint annotations to the sentence, for each possible display. The total expected disagreement for the sentence is the sum of these probabilities across all displays, using the same weights as above.

The overall observed disagreement in the corpus D_o is the arithmetic mean of the disagreement on each sentence; similarly, the overall expected disagreement D_e is the mean of the expected disagreement across all of the sentences. To compute the value of β for the output of the two coders, we subtract the ratio of these two values from 1:

$$\beta = 1 - \frac{D_o}{D_e}$$

As Artstein and Poesio (2005) point out, there is no significance test for agreement with weighted measures such as β, and the actual value is strongly affected by the distance metric that is selected. However, β values can be compared with one another to assess degrees of agreement. The overall β value between the two coders on the full set of 286 sentences processed by both was 0.561, with β values on individual facial displays ranging from a high of 0.661 on nodding to a low of 0.285 on eye narrowing (a very rare motion). To put these values into context, we also computed β on the set of 20 sentences processed by the additional coder as part of the training process (which are not included in the set of 286). The overall β value between the coders on these sentences is 0.231, with negative values for some of the individual displays, indicating that the training process had a positive effect on agreement.

3.4. Patterns in the corpus

Several contextual features had a significant effect on the facial displays occurring on that node. To determine the most significant factors, we performed multinomial logit regression as described by Fox (2002); the following contextual features had the most significant effect (all $p < 0.001$ on the Wald test). Nodding and brow raising were both

more frequent on nodes with any sort of predicted pitch accent. In negative user-preference contexts, eyebrow raising, eye narrowing, and left leaning were all relatively more frequent; in positive contexts, the relative frequency of right turns and brow raises was higher. In the first half of two-clause sentences, brow lowering was also more frequent, as was upward nodding, while downward nodding and right turns showed up more often in the second clause. The impact of all of these features was similar in the corpora produced by both annotators. Foster (2007) describes the corpus patterns in detail.

The increased frequency of nodding and brow raising in on prosodically accented words agrees with other findings on emphatic facial displays such as those of Ekman (1979) and Graf et al. (2002). The findings on characteristic positive and negative displays do not have any direct analog in previous work, but when these displays were shown to human judges, they were reliably able to identify them and preferred outputs with consistent polarity on the verbal and non-verbal channels (Foster, 2007). These findings from the corpus add to the growing body of knowledge on the communicative function of non-verbal signals: Krahmer and Swerts (2005), for example, have demonstrated that typical expressions of uncertainty are identifiable, while Rehm and André (2005) found that an embodied agent using deceptive non-verbal behaviour was seen as less trustworthy than one that did not.

3.5. Satisfying the requirements for a generation corpus

This corpus addresses all of the requirements for a generation corpus outlined in Section 2.2. As in many previous corpora, we ensured that the corpus included full contextual information by basing it on output created in known pragmatic contexts. Also like many others, we designed the annotation scheme to consider only those behaviours (head and eyebrow motions) that could easily be controlled on the talking head to be described in Section 5.2. Note in particular that we chose not to annotate the amplitude of mouth movements, despite the fact that it has been documented to be correlated with prosodic emphasis, because this is not a dimension that can easily be controlled on the target head.

In the final corpus, cross-modal links are made between facial displays and sets of nodes in the OpenCCG derivation tree, which is useful in the generation process and also allowed for respectable inter-coder agreement. Selecting a linking level took some effort and experimentation, and two other versions were considered before settling on the one in the final annotation scheme. We can use the data in the corpus to test whether these modifications to the scheme were justified.

In a previous study using the same video recordings but a different, simpler scheme (Foster and Oberlander, 2006), facial displays could only be associated with single leaf nodes (i.e., words); that is, in the terminology of Ekman (1979), all motions were considered to be *batons* rather than *underliners*. Based on the data in the current corpus, that restriction is clearly unrealistic: the mean number of nodes spanned by a display in the full corpus is 1.95, with a maximum of 15 and a standard deviation of 2. The results are similar in the sub-corpus produced by the additional coder, with a mean node span of 2.25.

Another extension to the original annotation scheme was to allow displays to be attached to more than one node in the tree in cases where the span of words was not a syntactic constituent. The corpus data also supports this extension: approximately 6% of the annotations in the main corpus—165 of 2826—were attached to more than one node in the derivation tree, while the additional coder attached 4.5% of displays to multiple nodes.

4. Generation strategies

Once the video had been annotated, we used the 444-sentence corpus produced by the primary annotator to select motions for a synthetic talking head. Based on the corpus analysis described in Section 3.4, we used the following node features to select facial displays: the user-preference evaluation, the clause, the pitch accent, and the surface string associated with the node with semantic-class replacement.

To choose displays for a sentence, we started with the labelled derivation tree created by the text planner (Figure 1). The algorithm then proceeded depth-first down the tree, choosing a set of displays for each node as it was encountered. For each node, we considered all corpus nodes with the same context and selected a display combination in one of two ways: taking the highest-probability option or making a weighted choice among all options.[2]

As a concrete example of the two generation strategies, consider a hypothetical context in which the speaker made no motion 80% of the time, a downward nod 10% of the time, and a nod with a brow raise the other 10% of the time. For a node with this context, the majority generation strategy would choose the majority option of no motion 100% of the time, while the weighted strategy would choose nothing with probability 0.8, a downward nod with probability 0.1,

[2] We did not select any motions on words for which the speech-synthesiser output was very short, such as *but* and *is*, because the synthesiser could not make those words long enough to make any motion sensible.

	Although	*it's*	*in*	*the*	*family*	*style,*	*the*	*tiles*	*are*	*by*	*Alessi Tiles.*
C	nd=d	nd=d	nd=d		nd=d				nd=d,bw=u		
	..... ln=l										
M					nd=d						nd=d
W	nd=d				nd=d		..tn=r..				

Figure 5. Face-display schedules for a sample sentence

and a nod with a brow raise with probability 0.1. Figure 5 shows the original corpus schedule (C) for the sentence in Figure 1, along with the schedules generated by the majority (M) and weighted (W) strategies.

5. Evaluation of generated output

We compared the schedules produced by the generation strategies in two ways. First, we used automated cross-validation to test how closely the strategies resembled the data in the corpus. We also performed a study in which human judges were asked to choose their preferred version among synthesised videos of the schedules generated by the two strategies and re-synthesised versions of the corpus examples.

5.1. Corpus-similarity evaluation

We compared the face-display schedules generated by the majority and weighted strategies through 10-fold cross-validation against the corpus, as follows. First, we divided the corpus at random into 10 equal-sized segments. For each segment, the counts of face-display combinations in each context were gathered using the other 90% of the corpus; these probabilities were then used to create display schedules for each of the sentences in the held-out 10%, using both of the generation strategies described in the preceding section.

We compared the generated schedules sentence-by-sentence against the facial displays found in the corpus, using a range of measures: precision, recall, F score, node accuracy, and β. For precision we counted the proportion of proposed motions that had exact matches in the corpus, while for recall, we counted the proportion of the corpus motions that were reproduced exactly in the generated output; the F score for a sentence was then the harmonic mean of these two values. Node accuracy reflects the proportion of nodes in the derivation tree where the proposed displays were correct, including those nodes where the algorithm correctly proposed no motion.[3] Overall scores were obtained by averaging the sentence-level scores across the corpus. We also computed

[3] A baseline system that never proposes any motion scores 0.79 on this measure.

Table I. Results for the corpus-similarity measures, averaged across sentences

	Majority					Weighted				
	Prec	*Rec*	*F*	*NAcc*	*Beta*	*Prec*	*Rec*	*F*	*NAcc*	*Beta*
Mean	0.52	0.31	0.18	0.82	0.34	0.29	0.24	0.12	0.75	0.23
Min	0.0	0.0	0.0	0.56	–	0.0	0.0	0.0	0.40	–
Max	1.0	1.0	0.5	1.0	–	1.0	1.0	0.4	0.95	–
Stdev	0.32	0.22	0.12	0.08	–	0.25	0.20	0.10	0.09	–

a value for β as described in Section 3.3 for each strategy, comparing the full set of generated sentences against the full set of corpus sentences.

Table I shows the results for all of these corpus-similarity measures, averaged across the sentences in the corpus. The majority strategy scored uniformly higher than the weighted strategy. The difference was particularly dramatic for precision, where the value for the majority strategy (0.52) was nearly twice that for the weighted strategy (0.29); that is, the motions proposed by majority strategy were identical to the corpus nearly twice as often. Using a T test, the differences on precision, recall, and node accuracy are all significant at $p < 0.001$; also, the node accuracy score for the majority strategy is significantly better than the no-motion baseline of 0.79, while that for the weighted strategy is significantly worse. Significance cannot be assessed for the differences in the F scores or β values, but the trend is the same.

5.2. Human preferences

The face-display schedules generated by the majority strategy scored above those generated by the weighted strategy on all corpus-similarity measures. However, the majority display combination in almost all contexts (88%) is actually no motion at all, and occasionally (7.1%) a downward nod on its own. This means that the schedules generated by the majority strategy tend to have nodding on accented words as the only motion type. These schedules score highly on corpus similarity because they do not diverge greatly on average from the corpus; however, this does not necessarily mean that such facial displays will be preferred by users over those generated by the weighted strategy, which include a wider range of the non-verbal behaviours recorded in the corpus. Indeed, in other studies of corpus-driven generation systems—e.g., Belz and Reiter (2006)—the versions preferred by human judges tended be those that scored lower on corpus similarity. To test whether that is the case with this system, we gathered human judgements on the generated output.

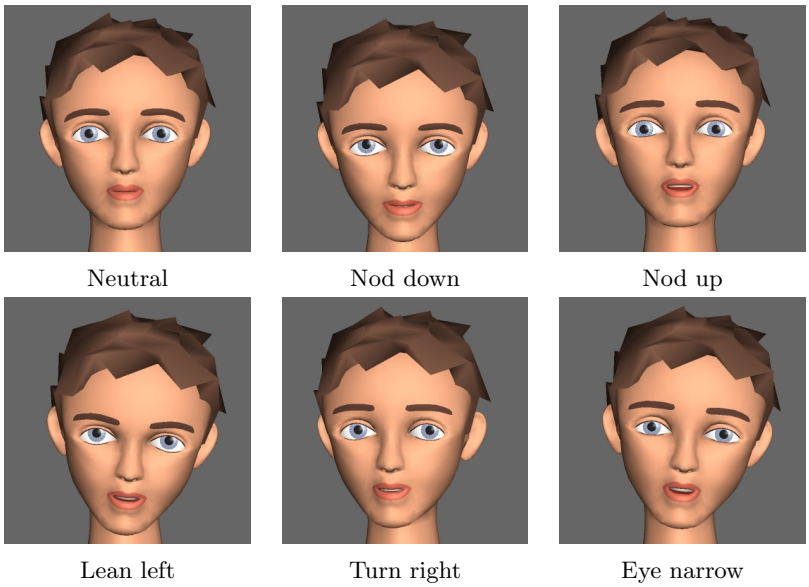

Figure 6. Synthesised facial expressions

5.2.1. *Materials*

We randomly selected 24 sentences from the corpus and generated three talking-head videos for each: using the schedules generated by the majority and weighted strategies in the cross-validation, as well as the original corpus annotations.[4] The videos were generated using the RUTH talking head (DeCarlo et al., 2004) and the Festival speech synthesiser (Clark et al., 2004), using built-in facial displays of the RUTH head synchronised with the relevant span in the speech. Figure 6 shows some sample facial displays on the RUTH head.

To map from a generation schedule to a RUTH video, we first obtained the phoneme timing for all words from Festival. We then created an animation schedule with the timing of all selected motions, where each motion in the schedule was synchronised with the start and end of the corresponding words. For example, the right turn in the weighted schedule from Figure 5 would begin with the first phoneme of *the* and end at the same time as the *s* of *tiles*. Every instance of the same motion (e.g., a downward nod) was realised with the same low-level RUTH commands: built-in commands for brow motions and eye narrowing, and "jogs" for the rigid head motion. The schedule was then sent to RUTH along with the speech-synthesiser waveform to create a video.

[4] The corpus schedules were modified to remove motions on short words such as *but* and *is*, for the reasons discussed in Section 4.

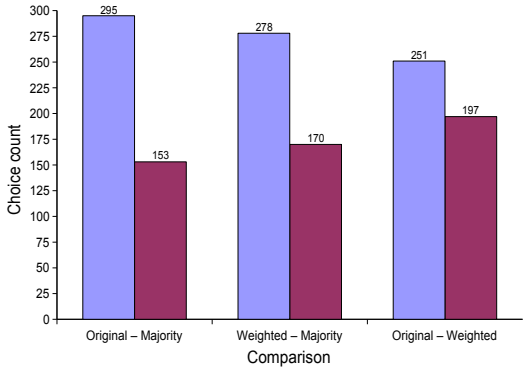

Figure 7. Overall human preference counts for each pairwise choice

5.2.2. *Procedure*

This experiment was run over the world-wide web, with subjects recruited through a department student mailing list and by a posting on a website devoted to psycholinguistic experiments. A total of 56 subjects took part: 34 male subjects and 22 females, mostly between the ages of 20 and 30. 31 of the subjects were expert computer users, while the rest were mainly intermediate users. Just under half of them (24) were native speakers of English, while most of the rest were speakers of other European languages.

Each subject was shown pairs of videos for all 24 sentences and asked to choose which version they preferred, following their first instinct as much as possible. Each subject performed each of the three possible pairwise comparisons between schedule types eight times, four times in each order. Both the mapping between pairwise comparisons and items and the presentation order were generated randomly for each subject.

5.2.3. *Results*

The overall results of this study are shown in Figure 7. Each pair of bars shows the count of pairwise choices made between schedule types; for example, when the choice was between an original corpus schedule and one generated by the majority strategy, the original version was chosen 295 of 448 times (66%). To assess the significance of the results, we can use a binomial test, which provides an exact measure of the statistical significance of deviations from a theoretically expected classification into two categories. This test indicates that all of the trends are significant: the original vs. weighted comparison at $p < 0.05$, and the other two at $p < 0.0001$. None of the demographic factors had any impact on these results.

5.3. Discussion

The results of the two studies differ: the cross-validation study scored the majority strategy higher on all measures, while the human subjects tended to prefer the output of the weighted strategy. This supports our prediction that the judges would prefer generated output that reproduced more of the variation in the corpus, regardless of the corpus-similarity scores; in this sense, these results are similar to those found by Belz and Reiter (2006).

The human judges preferred the regenerated corpus sentences to those generated by either strategy, although the preference over the weighted strategy was less pronounced. This suggests that making an independent choice for each node is useful, but not enough to capture the behaviour of the subject, and that more sophisticated generation strategies could be successful for this task. We discuss possible extensions in this area in the following section.

6. Conclusions and future work

We have presented a multimodal corpus based on a single speaker reading scripted sentences in the domain of the COMIC multimodal dialogue system, where the corpus was annotated for the head and eyebrow motions that occur in various syntactic, prosodic, and pragmatic contexts. The speaker showed systematic differences in the displays he used; the most relevant contextual factors were the user-preference evaluation, the predicted pitch accents, and the clause of the sentence. The characteristic behaviours on prosodically stressed words agree with previous findings on non-verbal behaviour; the motions correlated with positive and negative user-preference evaluations are more specific to this domain and corpus, but still sufficiently general that users were reliably able to identify them.

We used the data from this corpus to select head and eyebrow motions for an embodied conversational agent when producing output in this same domain. We compared two selection strategies: always choosing the majority option, or making a weighted choice among all of the options. The former strategy scored higher on every measure of corpus similarity in a cross-validation study, while the output of the latter strategy was preferred by human judges. This demonstrates the danger of relying on corpus similarity for evaluating generated output, as it tends to favour strategies that discard much of the interesting variation in the corpus. There is still a place for automated corpus-based evaluation in generation, particularly during the development of a

system or to verify that output is well-formed; however, it is crucial that any such evaluation be accompanied by a user study or an automated evaluation that considers other factors such as output diversity.

The human judges also preferred videos generated directly from the corpus data to the output of either strategy, with a more significant preference over the majority strategy. An interesting additional study would be to gather judgements on displays selected according to the overall corpus counts, independent of context. While the weighted strategy was partly successful, a more sophisticated implementation that better reproduces the range of corpus data would likely have greater success. COMIC uses the OpenCCG realiser, which incorporates n-gram models into its realisation process, so one possible implementation technique would be to build models combining words with multimodal behaviour and to replace the two-stage process by an integrated one. Such an implementation would also be more in line with the psycholinguistic evidence (McNeill, 2000) that verbal and non-verbal behaviour are produced together from a common representation.

Another possible source of increased output quality is to extend the range of displays. The annotation scheme for this corpus used only five motion types, and the RUTH videos were generated using a single example for each type, varying only in duration. For future implementations, a richer set of displays—gathered through motion capture or a different style of annotation—could produce more interesting and naturalistic output. To support such an implementation, the process of controlling the embodied agent would also have to be extended to support the full set of displays, and it is possible that supporting such displays would require a different embodied-agent implementation.

References

Artstein, R. and M. Poesio: 2005, 'Kappa^3 = alpha (or beta)'. Technical Report CSM-437, University of Essex Department of Computer Science.

Bangalore, S., O. Rambow, and S. Whittaker: 2000, 'Evaluation metrics for generation'. In: *Proceedings of INLG 2000*.

Belz, A., A. Gatt, E. Reiter, and J. Viethen: 2007, 'First NLG Shared Task and Evaluation Challenge on Attribute Selection for Referring Expression Generation'. `http://www.csd.abdn.ac.uk/research/evaluation/`.

Belz, A. and E. Reiter: 2006, 'Comparing Automatic and Human Evaluation of NLG Systems'. In: *Proceedings of EACL 2006*. pp. 313–320.

Belz, A. and S. Varges (eds.): 2005, 'Corpus Linguistics 2005 Workshop on Using Corpora for Natural Language Generation'.

Cassell, J., T. Bickmore, H. Vilhjálmsson, and H. Yan: 2001a, 'More than just a pretty face: Conversational protocols and the affordances of embodiment'. *Knowledge-Based Systems* **14**(1–2), 55–64.

Cassell, J., Y. Nakano, T. W. Bickmore, C. L. Sidner, and C. Rich: 2001b, 'Non-Verbal Cues for Discourse Structure'. In: *Proceedings of ACL 2001*.

Cassell, J., J. Sullivan, S. Prevost, and E. Churchill: 2000, *Embodied Conversational Agents*. MIT Press.

Clark, R. A. J., K. Richmond, and S. King: 2004, 'Festival 2 – build your own general purpose unit selection speech synthesiser'. In: *Proceedings of the 5th ISCA Workshop on Speech Synthesis*.

de Carolis, B., V. Carofiglio, and C. Pelachaud: 2002, 'From discourse plans to believable behavior generation'. In: *Proceedings of INLG 2002*.

DeCarlo, D., M. Stone, C. Revilla, and J. Venditti: 2004, 'Specifying and animating facial signals for discourse in embodied conversational agents'. *Computer Animation and Virtual Worlds* **15**(1), 27–38.

Ekman, P.: 1979, 'About brows: Emotional and conversational signals'. In: M. von Cranach, K. Foppa, W. Lepenies, and D. Ploog (eds.): *Human Ethology: Claims and limits of a new discipline*. Cambridge University Press.

Foster, M. E.: 2007, 'Evaluating the impact of variation in embodied output'. Ph.D. thesis, School of Informatics, University of Edinburgh.

Foster, M. E. and J. Oberlander: 2006, 'Data-driven generation of emphatic facial displays'. In: *Proceedings of EACL 2006*. pp. 353–360.

Foster, M. E., M. White, A. Setzer, and R. Catizone: 2005, 'Multimodal Generation in the COMIC Dialogue System'. In: *Proceedings of the ACL 2005 Demo Session*.

Fox, J.: 2002, *An R and S-Plus companion to applied regression*. Sage Publications.

Graf, H., E. Cosatto, V. Strom, and F. Huang: 2002, 'Visual Prosody: Facial Movements Accompanying Speech'. In: *Proceedings of FG 2002*. pp. 397–401.

Kipp, M.: 2004, *Gesture Generation by Imitation - From Human Behavior to Computer Character Animation*. Dissertation.com.

Krahmer, E. and M. Swerts: 2005, 'How children and adults produce and perceive uncertainty in audiovisual speech'. *Language and Speech* **48**(1), 29–53.

Langkilde, I. and K. Knight: 1998, 'Generation that Exploits Corpus-Based Statistical Knowledge'. In: *Proceedings of COLING-ACL 1998*.

Langkilde-Geary, I.: 2002, 'An Empirical Verification of Coverage and Correctness for a General-Purpose Sentence Generator.'. In: *Proceedings of INLG 2002*.

Mana, N. and F. Pianesi: 2006, 'HMM-based synthesis of emotional facial expressions during speech in synthetic talking heads'. In: *Proceedings of ICMI 2006*.

Martin, J.-C., P. Kühnlein, P. Paggio, R. Stiefelhagen, and F. Pianesi (eds.): 2006, 'LREC 2006 Workshop on Multimodal Corpora: From Multimodal Behaviour Theories to Usable Models'.

McNeill, D. (ed.): 2000, *Language and Gesture: Window into Thought and Action*. Cambridge University Press.

Passonneau, R. J.: 2004, 'Computing reliability for coreference annotation'. In: *Proceedings, Fourth International Conference on Language Resources and Evaluation (LREC 2004)*, Vol. 4. Lisbon, pp. 1503–1506.

Rehm, M. and E. André: 2005, 'Catch Me If You Can – Exploring Lying Agents in Social Settings'. In: *Proceedings of AAMAS 2005*. pp. 937–944.

Steedman, M.: 2000, 'Information structure and the syntax-phonology interface'. *Linguistic Inquiry* **31**(4), 649–689.

Stone, M., D. DeCarlo, I. Oh, C. Rodriguez, A. Lees, A. Stere, and C. Bregler: 2004, 'Speaking with hands: Creating animated conversational characters from recordings of human performance'. *ACM Trans. Graphics* **23**(3), 506–513.

White, M.: 2006, 'Efficient realization of coordinate structures in combinatory categorial grammar'. *Research on Language and Computation* **4**(1), 39–75.