Associating Facial Displays with Syntactic Constituents for Generation

Mary Ellen Foster
Informatik VI: Robotics and Embedded Systems
Technical University of Munich
Boltzmannstraße 3, 85748 Garching, Germany
foster@in.tum.de

Abstract

We present an annotated corpus of conversational facial displays designed to be used for generation. The corpus is based on a recording of a single speaker reading scripted output in the domain of the target generation system. The data in the corpus consists of the syntactic derivation tree of each sentence annotated with the full syntactic and pragmatic context, as well as the eye and eyebrow displays and rigid head motion used by the the speaker. The behaviours of the speaker show several contextual patterns, many of which agree with previous findings on conversational facial displays. The corpus data has been used in several studies exploring different strategies for selecting facial displays for a synthetic talking head.

1 Introduction

An increasing number of systems designed to automatically generate linguistic and multimodal output now make use of corpora to help in decision-making (cf. Belz and Varges, 2005). Some implementations use corpora to help select output that is grammatical or fluent; for example, Langkilde and Knight (1998) and White (2006) both used n-gram language models to guide stochastic surface realisers. In other systems, corpora are used to make decisions based on pragmatic factors such as the reading level of the target user (Williams and Reiter, 2005) or the visual features of an object being described (Cassell et al., 2007). The latter type of domain-specific contextual information is not often included in generally-available corpora. For this reason, developers of generation systems that need this type of information often create and make use of application-specific corpora.

The easiest method of including the necessary pragmatic information in a corpus is to base the corpus on output generated in situations where the contextual factors are known; this eliminates the need to annotate these factors explicitly. Stone et al. (2004), for example, created a multimodal corpus based on the voice and body language of an actor performing scripted output in the domain of the target generation system: an animated instructor character for a snowboarding video game. The contextual information in the corpus scripts included the move that the player attempted in the game and the result of that attempt. Similarly, van Deemter et al. (2006) created a corpus of multimodal referring expressions produced in specific pragmatic contexts and used it to compare several referring-expression generation algorithms to human performance.

In this work, the task is to select facial displays for an animated talking head to use while presenting output in the COMIC multimodal dialogue system (Foster et al., 2005), which generates spoken descriptions and comparisons of bathroom-tile options. The output of the COMIC text planner includes a range of information in addition to the text: the syntactic derivation tree, the user's evaluation of the object being described, the information status (new or old, contrastive) of each fact described, and the predicted speech-synthesiser prosody. All of this contextual information can be used to help select

Proceedings of the Linguistic Annotation Workshop, pages 25–32,
Prague, June 2007.

appropriate facial displays to accompany the spoken presentation; however—as in the other systems mentioned above—this requires a corpus where the full context for every facial display is known. To create such a corpus, we recorded a speaker performing scripted output in the domain of COMIC.

This paper is arranged as follows. In Section 2, we first describe how the scripts for the corpus were created and how the recording was made. Section 3 then presents the annotation scheme and the tool that was used to perform the annotation, while Section 4 describes the measures that were taken to ensure that the annotation was reliable. Section 5 then summarises the high-level patterns that were found in the displays annotated in the corpus and compares them to other findings on conversational facial displays. At the end of the section, we use the corpus data to test two assumptions that were made in the annotation scheme. After that, in Section 6, we describe several experiments in which different methods of using the data in this corpus to select facial displays for a synthetic head have been compared. Finally, in Section 7, we summarise the contributions of this paper and draw some conclusions about the usefulness of this corpus for its intended task.

2 Recording

For this corpus, we recorded a single speaker reading a set of 444 scripted sentences in the domain of the COMIC multimodal dialogue system. The sentences were generated by the full COMIC output-generation process, 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 annotated with all of the available syntactic and pragmatic information from the output planner, including the following features:

- The user-model 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;

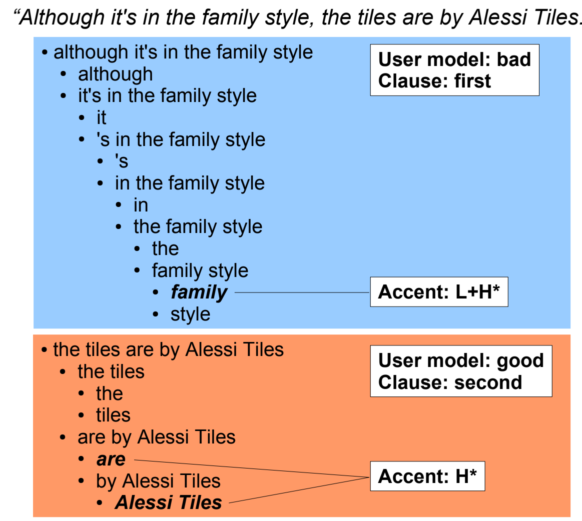


Figure 1: Annotated OpenCCG derivation tree

- 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 annotated OpenCCG derivation tree for a sample sentence drawn from the recording script. The annotations indicate that every node in the first half of this sentence is associated with a negative user-model evaluation and is in the first clause of a two-clause sentence, while every node in the second half is linked to a positive evaluation and is in the second clause of the sentence. The figure also shows the pitch accents selected by the output planner according to Steedman's (2000) theory of information structure and intonation.

For the recording, the sentences in the script were presented one at a time to the speaker; the presen-

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

tation included both the linguistic content (with accented words highlighted) as well as 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 Annotation

Once all of the sentences in the script had been recorded as described in the preceding section, the next step was to annotate the facial displays that occurred. We first used Anvil (Kipp, 2004) to split the video into individual clips corresponding to each sentence. This section describes how the facial displays in each of the clips were then annotated.

3.1 Annotation scheme

We annotated the speaker's facial displays by linking each to the span of nodes in the OpenCCG derivation tree with which it was temporally related. Making cross-modal links at this level made it possible to use the annotated information directly in the output-generation process for the experiments described in Section 6.

A display was associated with the full span of words that it coincided with temporally, as follows. If a single node in the derivation tree covered exactly all of the relevant words, then the annotation was placed on that node; if the words spanned by a display did not coincide with a single node, it was attached to the set of nodes that did span 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 temporally associated with that sequence would be attached to that node. On the other hand, if there were a motion associated with *the tiles are*, it would be attached to both the *the tiles* node and the *are* node.

The following were the features that were considered; for each feature, we note the corresponding Action Unit (AU) from the well-known Facial Action Coding System (Ekman et al., 2002).

- Eyebrows: up (AU 1+2) or down (AU 4)
- Eye squinting (AU 43)

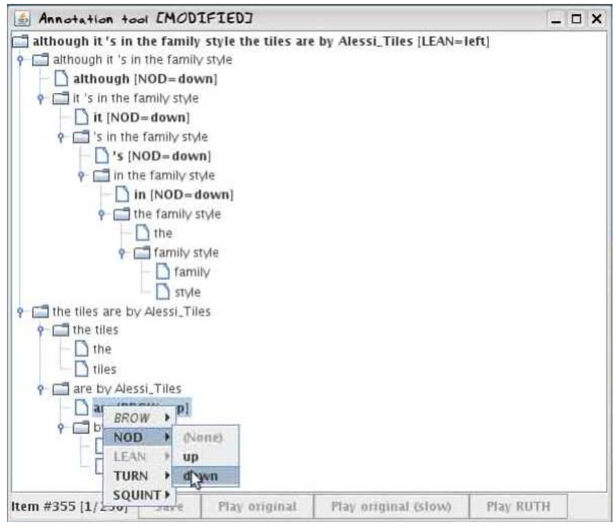


Figure 2: Annotation tool

- Head nodding: up (AU 53) or down (AU 54)
- Head leaning: left (AU 55) or right (AU 56)
- Head turning: left (AU 57) or right (AU 58)

This set of displays was chosen based on a combination of three factors: the emphatic facial displays documented in the literature, the capabilities of the target talking head, and the actual displays of the speaker during the recording session.

3.2 Annotation tool

The tool for the annotation was a custom-written program that 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 the coder to play back a proposed annotation sequence on a synthetic talking head to verify that it was as close as possible to the actual motions. Figure 2 shows a screenshot of the annotation tool in use on the sentence from Figure 1. In the screenshot, a left turn is 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 annotator has already attached a brow raise to the word *are* in the second half and is in the process of adding a nod to the same word.

The output of the annotation tool is an XML document including the original contextually-annotated

```
<node surf="although it 's in the family style the tiles are by Alessi_Tiles" LEAN="left"
    sc="although [pro3n] be in the [style] [abstraction] the [phys-obj] be by [manufacturer]">
  <node surf="although it 's in the family style" um="b" first="y"
      sc="although [pro3n] be in the [style] [abstraction]">
    <node surf="although" um="b" first="y" NOD="down" />
    <node surf="it 's in the family style" um="b" first="y"
        sc="[pro3n] be in the [style] [abstraction]">
      <node surf="it" stem="pro3n" um="b" first="y" NOD="down" />
      <node surf="'s in the family style" um="b" first="y" sc="be in the [style] [abstraction]">
        <node surf="'s" stem="be" um="b" first="y" NOD="down" />
        <node surf="in the family style" um="b" first="y" sc="in the [style] [abstraction]">
          <node surf="in" um="b" first="y" NOD="down" />
          <node surf="the family style" um="b" first="y" sc="the [style] [abstraction]">
            <node surf="the" um="b" first="y" />
            <node surf="family style" um="b" first="y" sc="[style] [abstraction]">
              <node surf="family" sc="[style]" accent="L+H*" um="b" first="y" NOD="down" />
              <node surf="style" sc="[abstraction]" um="b" first="y" />
            </node>
          </node>
        </node>
      </node>
    </node>
  </node>
  <node surf="the tiles are by Alessi_Tiles" um="g" first="n"
      sc="the [phys-obj] be by [manufacturer]">
    <node surf="the tiles" um="g" first="n" sc="the [phys-obj]">
      <node surf="the" um="g" first="n" />
      <node surf="tiles" sc="[phys-obj]" stem="tile" um="g" first="n" />
    </node>
    <node surf="are by Alessi_Tiles" um="g" first="n" sc="be by [manufacturer]">
      <node surf="are" stem="be" accent="H*" um="g" first="n" BROW="up" NOD="down" />
      <node surf="by Alessi_Tiles" um="g" first="n" sc="by [manufacturer]">
        <node surf="by" um="g" first="n" />
        <node surf="Alessi_Tiles" sc="[manufacturer]" accent="H*" um="g" first="n" />
      </node>
    </node>
  </node>
</node>
```

Figure 3: Annotated sentence from the corpus

OpenCCG derivation tree of each sentence, with each node additionally labelled with a (possibly empty) set of facial displays. Figure 3 shows the fully-annotated version of the sentence from Figure 1. This document includes the contextual features from the original tree, indicated by italics: every node in the first subtree has um="b" and first="y", while every node in the second subtree has um="g" and first="n", while the accented items also have an accent feature. Every node also specifies the string generated by the subtree that it spans, both in its surface form (surf) and with semantic-class and stem replacement (sc). This tree also includes the facial displays added by the coder in Figure 2, indicated by underlining: (LEAN="left") attached to the root node), a number of downward nods (NOD="down") on individual words in the first half of the sentence, and a nod accompanied by a brow raise (BROW="up") on *are* near the end.

4 Reliability of the 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 an initial annotation scheme. The outputs of these two coders were compared, and the coders discussed the differences and agreed on a revised scheme. One of these coders then used the final scheme to process the entire set of 444 sentences. As a further test of reliability, an additional coder was instructed on the use of the annotation tool and scheme and used them to process 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, it is necessary to define a measure that computes the distance between two proposed annotations. In this case, to compute the observed disagreement $D_o(S)$ on a sentence S, we use a measure similar to that proposed by Passonneau (2004) for measuring agreement on set-valued annotations. For each display proposed by each coder on the sentence, 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 proposed exactly the same display, 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 from the second coder, then that indicates the maximum level of disagreement (1). The total observed disagreement on a sentence is the sum of the disagreement level for each display proposed by each coder.

The expected disagreement $D_e(S)$ for a sentence S depends on the length of that sentence, as follows. 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 the observed disagreement 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:

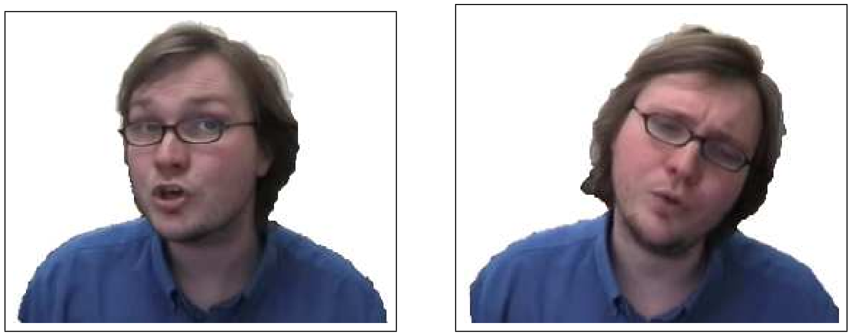
$$\beta = 1 - \frac{D_o}{D_e}$$

As Artstein and Poesio (2005) point out, for weighted measures such as β, there is no significance test for agreement, 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 squinting (a very rare motion). To put these values into context, we computed β on the set of 20 sentences processed by the final coder as part of the training process (which are not included in the set of 286). The overall β value for these sentences is 0.231, with negative values for some of the individual displays. This demonstrates that the training process had a positive effect on agreement.

5 Patterns in the corpus

We investigated the contextual features to see which had the most significant effect on the facial displays occurring on a node. To determine this, we used multinomial logit regression to select the factors and factor interactions that had the most significant effects on the distribution of each display; this form of regression is appropriate when, as in this case, the response variable is categorical. In this section, we list the most significant factors and give a qualitative description of the impact of each.

The single most influential contextual factor was the user-model evaluation, which had an effect on all of the facial displays. In positive user-model contexts, eyebrow raising and turning to the right were relatively more frequent (Figure 4(a)); in negative contexts, on the other hand, the rates of eyebrow lowering, squinting, and leaning to the left were all higher (Figure 4(b)). Other factors also affected the



(a) Positive (b) Negative

Figure 4: Characteristic facial displays for different user-model evaluations

distribution of facial displays. 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 of two-clause sentences. Nodding and brow raising were both more frequent on nodes with any sort of predicted pitch accent.

Several of these factors agree with previous findings on conversational body language. The increased frequency of nodding and brow raising on accented words agrees with many previous studies: Ekman (1979), Cavé et al. (1996), Graf et al. (2002), Keating et al. (2003), Krahmer and Swerts (2004), and Flecha-García (2006) all noted similar displays on prosodically accented parts of the sentence. The speaker's tendency to move right on positive descriptions and left on negative descriptions is also consistent with other findings. According to the work of Davidson and colleagues (Davidson and Irwin, 1999), emotion and affect processing are asymmetrically organised in the human brain. The right hemisphere is associated with negative affect (and withdrawal behaviours), and the left with positive affect (and approach behaviours). Because both perceptual and motor systems are contra-laterally organised, this means that higher levels of right hemisphere activity are associated with attention being oriented towards the left, while higher levels of left hemisphere activity are associated with attention being oriented to the right; this fits with our speaker's pattern of movements.

The annotation scheme described here allowed a display to be associated with any contiguous span of words in the sentence. Annotators were encouraged to use syntactic constituents wherever possible, but also had the option to select multiple nodes where a display did not correspond with a single constituent in the derivation tree. Earlier versions of the annotation scheme did not support this degree of flexibility, so we used the patterns in the corpus to test whether the modifications to the scheme were useful.

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 was clearly unrealistic: the mean number of nodes spanned by a display in the full corpus was 1.95, with a maximum of 15 and a standard deviation of 2. The results were similar in the sub-corpus produced by the final coder, in which the mean number of nodes spanned by a display was 2.25.

The annotation rules for this study did not initially permit displays to be associated with more than nodes in the derivation tree. This capability was added following inter-coder discussions after the initial test annotation to deal with cases where the speaker's displays did not correspond to syntactic constituents—for example, if the speaker raised his eyebrows on *the tiles are* or some other such non-standard constituent. The data in the annotated corpus supports this modification. Approximately 6% of the annotations in the main corpus—165 of 2826—were attached to more than one node in the derivation tree; for the final coder, 4.5% of annotations were on multiple nodes.

6 Generation experiments

The primary reason for creating this corpus of facial displays was to use the resulting data to select facial displays for the artificial talking head in the COMIC multimodal dialogue system. Several different strategies have been implemented to use the corpus data for this task, and a number of automated and human evaluations have been carried out comparing the different implementations.

As described in the preceding section, the factor with the largest influence on the displays of the recorded speaker was the user-model evaluation. Two studies (Foster, 2007b) were carried out to test the generality of the characteristic positive and negative displays (Figure 4). In the first study, users were asked to identify the intended user-model polarity of a description presented by the talking head based only on the facial displays. The participants were generally able to recognise the characteristic positive and negative facial displays; they also identified the displays intended to be neutral (nodding alone) as positive, and tended to judge videos with no facial displays to be negative. In the second study, users' subjective preferences were gathered between videos in which the user-model evaluation expressed in speech was either consistent or inconsistent with the facial displays. In this study, the participants generally preferred the videos that showed consistent content on the two output channels.

In another study (Foster and Oberlander, 2007), two different data-driven strategies were implemented that used the corpus data to select facial displays to accompany speech. One strategy always selected the highest-probability option in all contexts, while the other made a stochastic choice among all of the options weighted by the corpus probabilities. These two strategies were compared against each other using both automated and human evaluation methods: the majority strategy scored more highly on the automated cross-validation, while the weighted strategy was strongly preferred by human judges. The judges also preferred resynthesised versions of the original facial displays from the corpus to the output of either of the generation strategies.

Two further human evaluation studies compared the weighted data-driven generation strategy from the preceding study to a rule-based strategy that selected the most characteristic displays based only on the user-model evaluation (Foster, 2007a). When users' subjective judgements were gathered as above, they had a mild preference for the output of the weighted strategy over that of the rule-based strategy. In a second study, videos generated by the weighted strategy significantly decreased participants' ability to select descriptions that were correctly tailored to a given set of user preferences, while videos generated by the rule-based strategy had no such impact.

7 Conclusions

We have described the collection and annotation of an application-specific corpus of conversational facial displays. The designs of both the corpus and the annotation scheme were driven by the needs of a specific generation system, which makes use of a range of pragmatic information while creating output. To use this information to make corpus-based decisions, it is necessary that the full context of every utterance and facial display in the corpus be available. Rather than adding this information to an existing corpus, we chose—like Stone et al. (2004) and van Deemter et al. (2006), for example—to create a corpus based on known contexts so that the full information for every sentence was known before the fact.

The final annotation scheme required each facial display to be linked to the set of nodes in the syntactic derivation tree of the sentence that exactly covered the words temporally associated with the display. Two coders separately processed the sentences in the corpus; on the sentences processed by both coders (about 65% of the corpus), the agreement as measured by β was 0.561.

A number of contextual factors had an influence on the displays used by the recorded speaker. The single most influential factor was the user-model evaluation of the object being described. The speaker's characteristic side-to-side motions on these sentences agree with findings on the relationship between brain hemispheres and affect. In addition, in user studies, human judges were reliably able to identify the intended affect based on resynthesised versions of these characteristic displays. Other patterns in the data also agree with existing findings on facial displays: for example, the speaker tended to nod and raise his eyebrows more frequently on words with prosodic accents.

Several experiments have been performed in which the annotated data from this corpus was used to select the facial displays to accompany the output of an animated talking head. These studies have found interesting results on both the relationship between automated and human judgements of output quality and the relative utility of rule-based and data-driven approaches for selecting conversational facial displays.

Acknowledgements

This research was supported by the EU projects COMIC (IST-2001-32311) and JAST (FP6-003747-IP). Thanks to Amy Isard, Ron Petrick, and Tom Segler for annotation assistance, and to Jon Oberlander and the LAW reviewers for useful comments.

References

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

A. Belz and S. Varges, editors. 2005. *Corpus Linguistics 2005 Workshop on Using Corpora for Natural Language Generation*. http://www.itri.brighton.ac.uk/ucnlg/ucnlg05/.

J. Cassell, S. Kopp, P. Tepper, K. Ferriman, and K. Striegnitz. 2007. Trading spaces: How humans and humanoids use speech and gesture to give directions. In T. Nishida, editor, *Engineering Approaches to Conversational Informatics*. Wiley. In press.

C. Cavé, I. Guaïtella, R. Bertrand, S. Santi, F. Harlay, and R. Espesser. 1996. About the relationship between eyebrow movements and F_0 variations. In *Proceedings of the 4th International Conference on Spoken Language Processing (ICSLP 1996)*.

R. J. Davidson and W. Irwin. 1999. The functional neuroanatomy of emotion and affective style. *Trends in Cognitive Sciences*, 3(1):11–21. doi:10.1016/S1364-6613(98)01265-0.

K. van Deemter, I. van der Sluis, and A. Gatt. 2006. Building a semantically transparent corpus for the generation of referring expressions. In *Proceedings of the Fourth International Natural Language Generation Conference*, pages 130–132. Sydney, Australia. ACL Anthology W06-1420.

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

P. Ekman, W. V. Friesen, and J. C. Hager. 2002. *Facial Action Coding System*. A Human Face, Salt Lake City.

M. L. Flecha-García. 2006. *Eyebrow raising in dialogue: Discourse structure, utterance function, and pitch accents*. Ph.D. thesis, Department of Theoretical and Applied Linguistics, University of Edinburgh.

M. E. Foster. 2007a. Comparing rule-based and data-driven selection of facial displays. In *Proceedings of the ACL 2007 Workshop on Embodied Language Processing*.

M. E. Foster. 2007b. Generating embodied descriptions tailored to user preferences. In submission.

M. E. Foster and J. Oberlander. 2006. Data-driven generation of emphatic facial displays. In *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2006)*, pages 353–360. Trento, Italy. ACL Anthology E06-1045.

M. E. Foster and J. Oberlander. 2007. Corpus-based generation of conversational facial displays. In submission.

M. E. Foster, M. White, A. Setzer, and R. Catizone. 2005. Multimodal generation in the COMIC dialogue system. In *Proceedings of the ACL 2005 Demo Session*. ACL Anthology W06-1403.

H. Graf, E. Cosatto, V. Strom, and F. Huang. 2002. Visual prosody: Facial movements accompanying speech. In *Proceedings of the 5th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2002)*, pages 397–401. doi:10.1109/AFGR.2002.1004186.

P. Keating, M. Baroni, S. Mattys, R. Scarborough, and A. Alwan. 2003. Optical phonetics and visual perception of lexical and phrasal stress in English. In *Proceedings of the 15th International Congress of Phonetic Sciences (ICPhS)*, pages 2071–2074.

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

E. Krahmer and M. Swerts. 2004. More about brows: A cross-linguistic study via analysis-by-synthesis. In C. Pelachaud and Z. Ruttkay, editors, *From Brows to Trust: Evaluating Embodied Conversational Agents*, pages 191–216. Kluwer. doi:10.1007/1-4020-2730-3_7.

I. Langkilde and K. Knight. 1998. The practical value of *n*-grams in generation. In *Proceedings of the 9th International Natural Language Generation Workshop (INLG 1998)*. ACL Anthology W98-1426.

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

M. Steedman. 2000. Information structure and the syntax-phonology interface. *Linguistic Inquiry*, 31(4):649–689. doi:10.1162/002438900554505.

M. Stone, 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 Transactions on Graphics (TOG)*, 23(3):506–513. doi:10.1145/1015706.1015753.

M. White. 2006. Efficient realization of coordinate structures in Combinatory Categorial Grammar. *Research on Language and Computation*, 4(1):39–75. doi:10.1007/s11168-006-9010-2.

S. Williams and E. Reiter. 2005. Deriving content selection rules from a corpus of non-naturally occurring documents for a novel NLG application. In Belz and Varges (2005).