Generating output in the COMIC multimodal dialogue system

Mary Ellen Foster
School of Informatics
University of Edinburgh

W3C MMI Workshop
Sophia Antipolis, 20 July 2004

Overview

- The COMIC project and demonstrator
- Planning and generating output in COMIC
  - Multimodal fission in COMIC
  - Planning text, gestures, and facial expressions
  - Speech synthesis and output coordination
- System evaluation
- Next steps for fission

COMIC: "COnversational Multimodal Interaction with Computers"

- EU FP5 project: March 2002-Feb 2005
- Goal: apply results and models from cognitive psychology to multimodal dialogue
- Demonstrator: adds a multimodal dialogue interface to a CAD-like system for bathroom design
  - 1. Specify shape of bathroom
  - *(2. Place furniture)*
  - **3. Browse available tiles**

Input processing and dialogue management

- Speech recognition and NLP
- Handwriting and (pen-)gesture recognition
- Multimodal fusion
- Dialogue manager
- Dialogue history manager, ontology manager

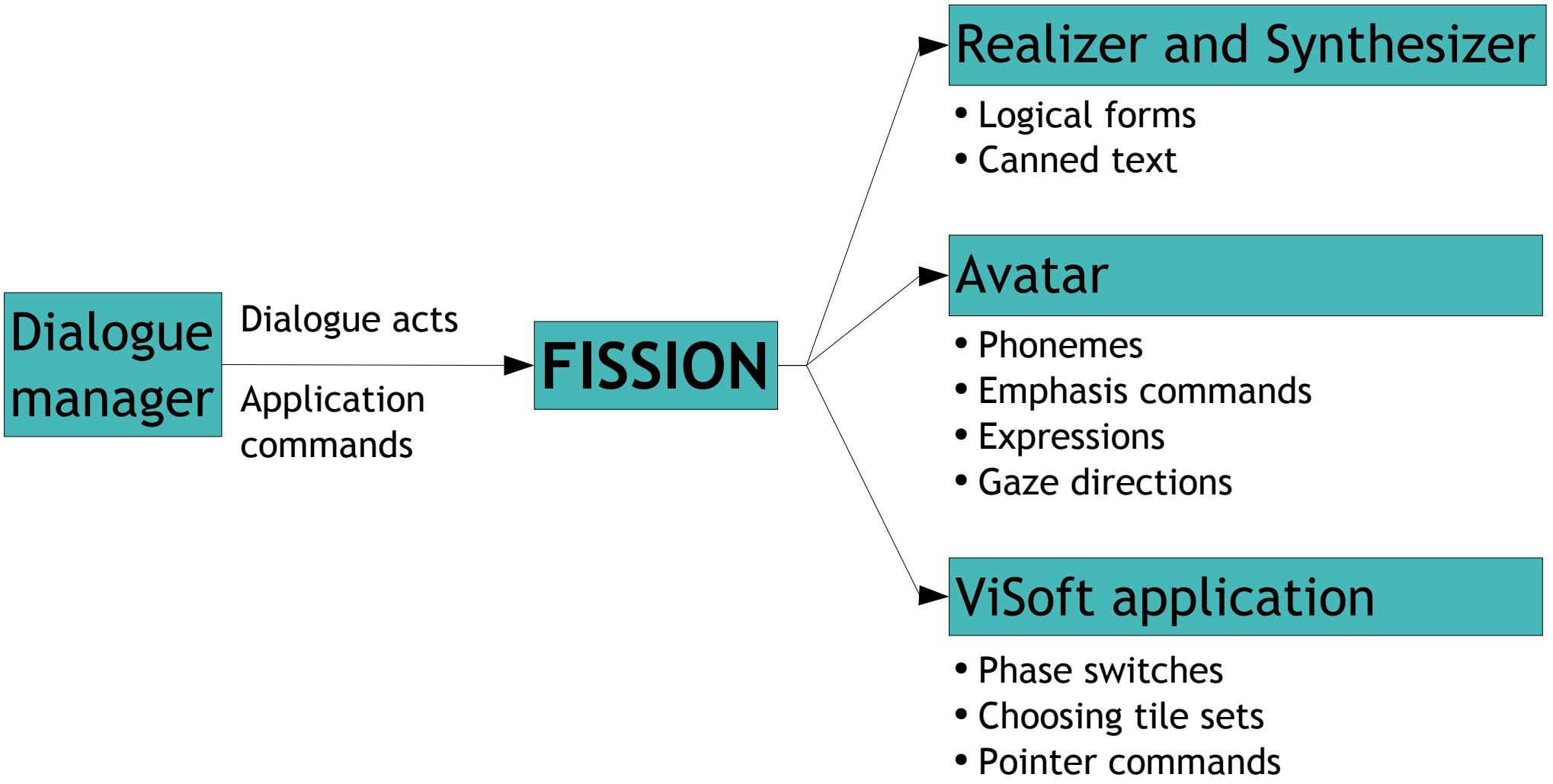
Fission and output processing

- **Fission module (presentation planner)**
- Speech synthesis (Edinburgh)
  - Surface realiser: OpenCCG (White, 2004)
  - Speech synthesiser: Festival, unit selection
- “Talking head” avatar
- Bathroom-design application

Sample interaction (browsing tiles)

- COMIC: [Introduction] ... “Are you ready?”
- User: “Yes.”
- COMIC: [Describes tiles on screen] ... “Please choose one.”
- User: “Show me this one.” [Circles second design]
- COMIC: [Chooses and describes tiles] ... “Do you want to see more modern designs?”
- ... *etc.* ...

Fission inputs and outputs

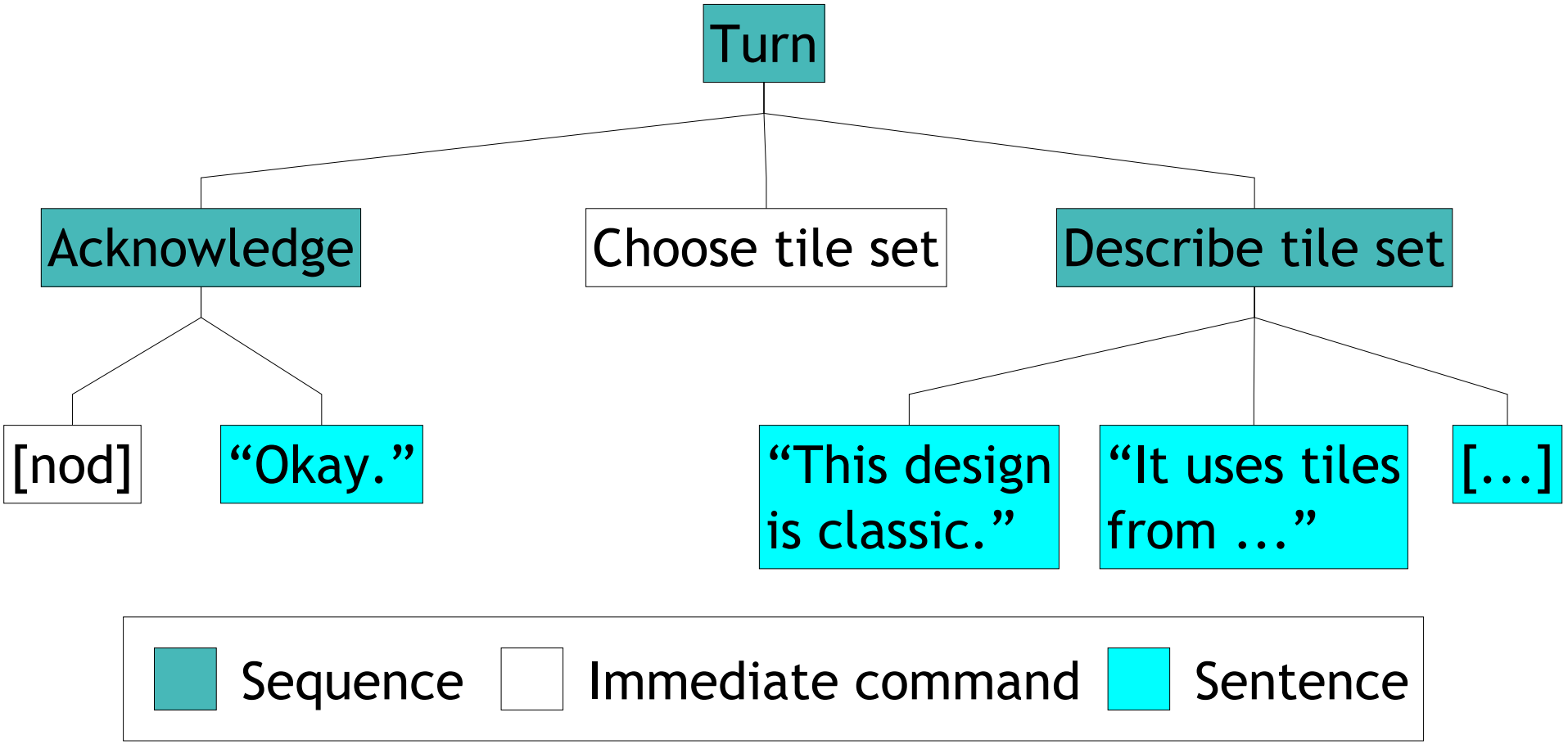


Fission tasks

- Content selection and structuring
  - Elaborate the high-level specification from the dialogue manager
- Modality selection
  - Decide on the content to be produced on each channel
- Output coordination
  - Ensure the output is coordinated temporally and spatially

Sample output plan

- **DAM input:** show(tileset21), describe(tileset21)



Creating and executing an output plan

- Create initial high-level structure based on DAM specification
- Elaborate and then output children in order
- Planning and execution are interleaved; later children in preparation while output is being produced from earlier ones
  - Avoid adding to (already non-trivial) latency

Text planning with XSLT (non-canned text)

- Gather information from system ontology; filter based on dialogue history; put in order
- Combine adjacent messages when possible
- Create a logical form (with alternatives) for each message and send it to the realiser
- Details:
  - M E Foster and M White. *Techniques for text planning with XSLT*. NLPXML-4 Workshop, 25 July 2004, Barcelona.

Speech synthesis

- Voice: general-purpose unit selection, with in-domain recording scripts
- Realiser output includes intonation, but current voice can't support it (stay tuned!)

```
<?xml version="1.0" encoding="UTF-8"?>
<!DOCTYPE apml SYSTEM "apml.dtd">
<apml>
  <performative>
    <emphasis x-pitchaccent="Hstar">This</emphasis>
    <emphasis x-pitchaccent="Hstar">design</emphasis>
    is
    <emphasis x-pitchaccent="Hstar">classic</emphasis>
    <boundary type="LL"/> .
  </performative>
</apml>
```

Speech timing

- Speech timing determines presentation timing
- Coordination achieved by adding labelled spans to the input of the speech module

```
<seg id="123">
  <speech>
    Hello
      <span label="ww">
        world
      </span>
    .
  </speech>
</seg>
```

```
<speech id="123">
  <words>
     <word id="w0" start="0.018750" end="0.334000" content="Hello">
       <phoneme id="p0" start="0.018750" end="0.101750" content="h"/>
       <phoneme id="p1" start="0.101750" end="0.114000" content="@"/>
       <phoneme id="p2" start="0.114000" end="0.194563" content="l"/>
       <phoneme id="p3" start="0.194563" end="0.334000" content="ou"/>
    </word>
    <word id="w1" start="0.334000" end="0.819688" content="world">
       <phoneme id="p4" start="0.334000" end="0.445750" content="w"/>
       <phoneme id="p5" start="0.445750" end="0.511813" content="@@r"/>
       <phoneme id="p6" start="0.511813" end="0.577188" content="r"/>
       <phoneme id="p7" start="0.577188" end="0.730187" content="l"/>
       <phoneme id="p8" start="0.730187" end="0.819688" content="d"/>
    </word>
  </words>
  <spans>
    <span type="labelled" info="ww" start="w1" end="w1"/>
  </spans>
</speech>
```

Planning pointer “gestures”

- Mark NPs in input with on-screen referents, and choose gestures and offsets for some subset
- Use application screen state to find objects
- Two versions: rule-based, or corpus-based
  - Evaluation (just completed): forced choice between two versions; justify choice where possible
- Details:
  - M E Foster. *Corpus-based planning of deictic gestures in COMIC*. INLG-04 (Student Session), Brockenhurst, 14-16 July 2004.

Facial expressions, gaze, and emphasis

- Expressions and gaze: only between sentences
- Phonemes: extracted from speech-synthesiser timing
- Emphasis commands: based on pitch accents

Output sequencing and coordination

- Sequences: Traverse subtree in order, waiting for any nodes that are not ready yet
- Immediate commands (expressions, gaze, screen-state changes): send command, wait for “done” report
- Sentences:
  - Send text to synthesiser (canned or via realiser)
  - Send timing to avatar; prepare gestures
  - Send “go at time *t*” + concrete gesture schedule

System evaluation

- Subjects use system for 15-20 minutes
  - Conditions: full face or “zombie”
- Measures
  - Recall of information presented (task success)
  - Subjective user-satisfaction questionnaire
  - Objective measures from log files
- Just completed (37 subjects); no results yet
- Evaluation of room-drawing phase pending

Next steps for fission

- Incorporate ideas from centering theory into text planning (Kibble & Power, 2000; Karamanis, 2003)
- Refer to a user model throughout the generation process (Moore et al., 2004)
- Holy grail: *instance-based* multimodal generation
  - Gather good instances by having users rate various combinations (as in current gesture evaluation)
  - Use (upcoming) factored language models in OpenCCG to choose among cross-modal alternatives

W3C standards

- Currently in use
  - XSLT, XPath: for text planning (NLPXML paper), plus *many* other stylesheets used internally
- Possible additions
  - SMIL: not for serialisation; possibly for internal data structures
  - SSML: if the synthesiser supports it
  - EMMA for output? Find out more
  - *(EMMA for input? can't comment)*

References

http://www.hcrc.ed.ac.uk/comic/

http://www.iccs.inf.ed.ac.uk/~mef/