

The *SACTI-2* Corpus:
Guide for Research Users

Karl Weilhammer, Jason D. Williams, Steve Young
{kw278, jdww30, sjy}@eng.cam.ac.uk

Technical Report: CUED/F-INFENG/TR.505

30 October 2004 (Version 0.1)

University of Cambridge
Department of Engineering

Table of contents

INTRODUCTION.....	4
AUDIENCE AND SCOPE.....	4
CORPUS OVERVIEW.....	4
DIFFERENCES BETWEEN SACTI-1 AND SACTI-2.....	5
ACKNOWLEDGEMENTS.....	5
CORPUS LICENSE.....	5
DATA COLLECTION ENVIRONMENT.....	5
OVERVIEW OF THE SIMULATED ASR-CHANNEL.....	5
TURN-TAKING MODEL IN THE SIMULATED ASR CHANNEL.....	7
THE CLICK MAP INTERFACE.....	7
DATA COLLECTION METHOD.....	8
CORPUS CONTENTS.....	9
SUMMARY.....	9
FILE NAMING AND ORGANIZATION.....	10
TRANSCRIPTIONS & ANNOTATION.....	11
USER UTTERANCE TRANSCRIPTION IN THE SIMULATED ASR CHANNEL.....	11
WIZARD UTTERANCE TRANSCRIPTION IN THE SIMULATED ASR CHANNEL.....	11
AUTOMATIC TURN ANNOTATION IN THE SIMULATED ASR DIALOGUES.....	11
TASK-COMPLETION ANNOTATION.....	11
DIALOGUE DETAILS.....	12
TEMPORAL RELATIONSHIP BETWEEN DIFFERENT TIERS.....	12
“states” track.....	13
“turns” track.....	14
“userUttsEP”, “wizardUttsEP” and “wizardUtts” track.....	14
“userClicks” track and “wizardClicks” track.....	14
“wizardButtons” track.....	15
DIALOGUE PROPERTIES SET.....	15
EXAMPLE SIMULATED ASR CHANNEL ANVIL FILE.....	16
REFERENCES.....	18
APPENDIX A: EXPERIMENTATION DOCUMENTS.....	19
APPENDIX B: WIZARD TRANSCRIPTION GUIDELINES.....	67
APPENDIX C: CONVERTING FROM HTK WAV TO RIFF WAV.....	70

Introduction

This document describes the collection procedure and transcription guidelines used for the collection of the *SACTI-2* dialogue corpus. It builds on previous work that was done collecting the *SACTI-1* corpus. *SACTI* stands for *Simulated ASR Channel, Tourist Information*. Both corpora can be used for training statistical components of speech only or multi-modal dialogue systems.

Corpus license

The corpus described in this document may be covered by a license agreement. Please see the corpus distribution itself for more information.

Audience and scope

This document is intended as a stand-alone “manual” for researchers interested in using the *SACTI-2* corpus for research endeavours. Large parts of this document follow texts of the manual for the *SACTI-1* corpus [13].

Previous Work

All previous work is related to the *SACTI-1* corpus. A discussion of the “Simulated ASR Channel” collection framework, and an analysis of the corpus for various conversational phenomena is presented in [2]. Further work is on-going. For details on the motivation of the collection framework, especially concerning the ASR confusion simulation, see [1].

Corpus overview and comparison of SACTI-1 and SACTI-2

The *SACTI* corpus consists of task-oriented dialogues between two people (volunteers). The two participants are called the “user” and the “wizard”. Each wizard was given a host of information about a fictitious town; each user is given a series of tasks to attempt to complete. The corpus is divided into 2 parts, *SACTI-1* and *SACTI-2*.

The *SACTI-1* corpus contains human-human dialogues. The major part of them was recorded in a “simulated automated speech recognition (ASR) channel” and a small portion was recorded as direct conversation. The participants could only communicate via speech. The simulated speech recognition error rate was varied during the recordings from no errors to high. Apart from the logging information of the recording setup, orthographic transcriptions, grounding acts and the understanding status of the wizard have been annotated. For further information on *SACTI-1*, please consult the respective Manual [13].

The *SACTI-2* corpus contains human-human dialogues in a “simulated automated speech recognition (ASR) channel.” The interaction is via speech and an interactive map. A lot of effort has been put in getting a good time resolution for mouse clicks with respect to the speech recordings. The user interface was slightly varied during the recordings. User and wizard behaviour responded to these variations. The simulated speech recognition error rate was no errors for one recording day and medium for all other recording days. Up to now, the logging information of the recording setup and orthographic transcriptions are available with the corpus. The data of *SACTI-1* is intended to be useful for investigating how speech and mouse clicks on a map can be used together in a conversation. The main differences between *SACTI-1* and *SACTI-2* are summarised in Table 1.

	SACTI-1	SACTI-2
Communication	speech	speech and interactive map
Time resolution	1 sec	0,001 - 0,5 sec depending on what was recorded
speech files	speech segments/turns	speech segments/turns full dialogue recordings for wizard and user
Different Tasks	24 standard tasks	24 standard tasks + 6 new tasks
Error rates	Non, Low, Med, Hi (all standard tasks)	Non (all tasks) Med (standard tasks), Hi (new tasks)
Users per wizard	3 users	6 users
Corpus contents	states sequence turns - - - - - - wizard understanding status grounding acts	states sequence turns user clicks wizard clicks wizard buttons user endpointed turns wizard endpointed turns manually transcribed and segmented wizard turns - -

Table 2: Comparison of SACTI-1 and SACTI-2

Acknowledgements

The authors would like to thank the following individuals for their support in this work:

1. Michael Kipp, for making his ANVIL tool available.
2. Carol Lions at Human Resources, Cambridge, for staffing wizards.
3. Jost Schatzmann, for helpful discussions regarding transcription conventions and tools.
4. Matt Stuttle for assistance with conducting the tests

This work was supported by the European Union Framework 6 TALK Project (507802).

Data collection environment

This section describes the simulated ASR channel, the multi-modal interface and the data collected. The framework is based on a “Wizard of Oz” trial but has been modified as summarised in Figure 1. Two experimental participants, the “subject” and the “wizard” communicate via a simulated ASR channel and an interactive map interface.

Overview of the simulated ASR-channel¹

Our methodology is similar to previous approaches [3], but introduces several additional elements of control.

The participants are located in different rooms and cannot see each other. It is also important that the participants do not interact until the test is concluded.

The subject can hear the wizard directly. However, the wizard cannot hear the subject; rather, both participants are told that the subject is speaking to a speech recogniser, which will take its best guess of what the subject says, and display it on a screen in front of the wizard.

¹ This section has been largely taken from [1].

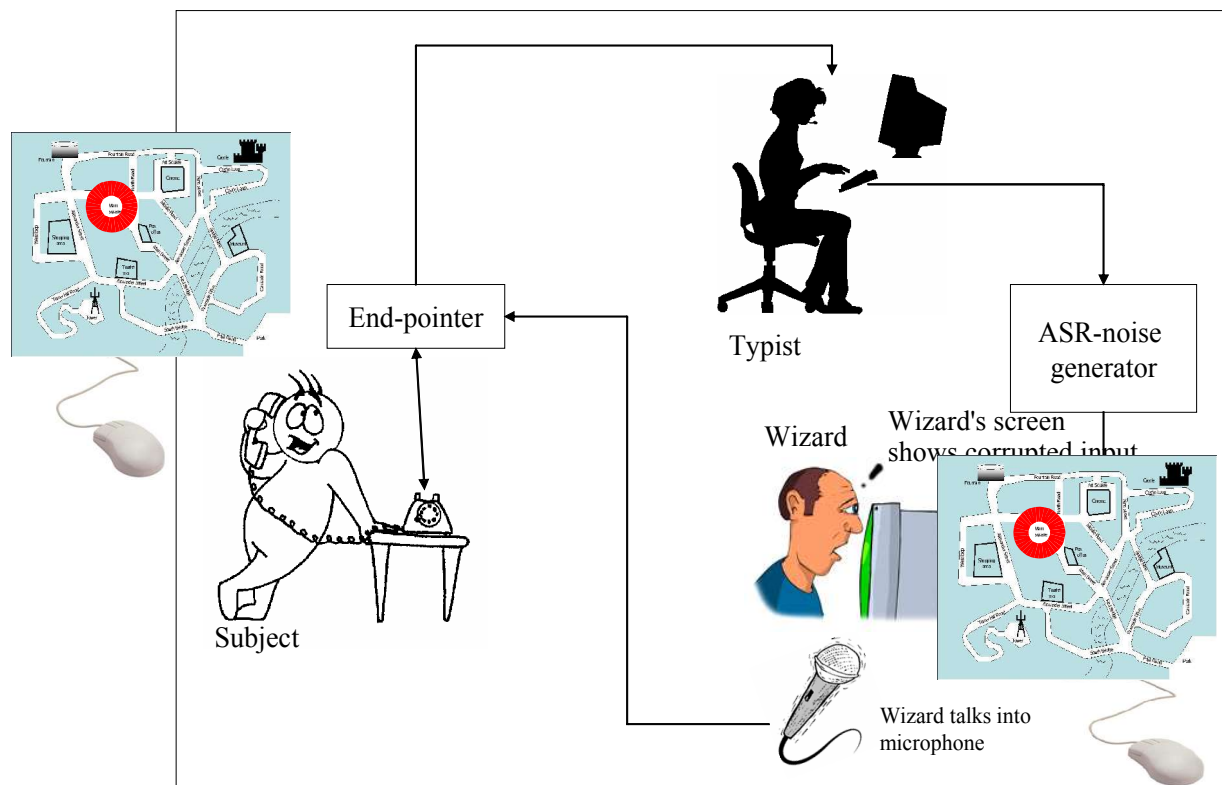


Figure 1: Schematic view of collection framework

When the system is busy and not listening to a participant, they hear a “tick-tock” sound. A turn-taking model patterned after typical human-computer turn-taking models is used in which the user may “barge-in over” (interrupt) the wizard, but the wizard may not interrupt the user.

In reality, the subject is speaking to a typist, who quickly transcribes the user’s utterance. This transcription is passed to a system which simulates ASR errors, the output of which is displayed to the wizard. The ASR simulation is used to control the WER, so that a variety of operating conditions can be explored and the collection can be initiated quickly.

The speech of both participants is end-pointed (i.e., segmented into utterances for performing recognition) using a standard energy-based end-pointer. The end-pointer is used to determine what wizard speech to play to the user, and what user speech to play to the typist. The end-pointing happens in just under real-time. The end-pointed utterances are saved for future analysis. One process maintains the state of the system and writes one system log.

The typist was asked to type the user’s speech as quickly and accurately as possible. The specific guidelines given to the typist are shown in *Form R-T* (see Appendix A).

The recording were done using sound blaster sound cards and KOSS CS-100 headsets.

Turn-taking model in the simulated ASR channel

Internally there are 5 system states:

1. **SILENCE:** Either participant can begin speaking; both hear silence.
2. **WIZARD TALKING:** Entered when the wizard starts talking. The user hears the wizard in this state. If the user interrupts, transition to USER TALKING: If the wizard stops speaking, transition to SILENCE.

3. **USER TALKING:** Entered when the user starts talking. The wizard hears the tick-tock sound, and the typist hears the user, and can begin typing. When the user finishes, transition to TYPIST TYPING.
4. **TYPIST TYPING:** Both participants hear the tick-tock sound. The typist can press a button to hear the user's utterance again. When the typist finishes typing, transition to CONFUSER CONFUSING.
5. **CONFUSER CONFUSING:** Both participants hear the tick-tock sound. The ASR error generator (the "confuser") receives the typist's text and produces the "confused" version, which is displayed on the wizard's screen. Once finished, transition to SILENCE.

Transitions between states are summarised in Figure 2.

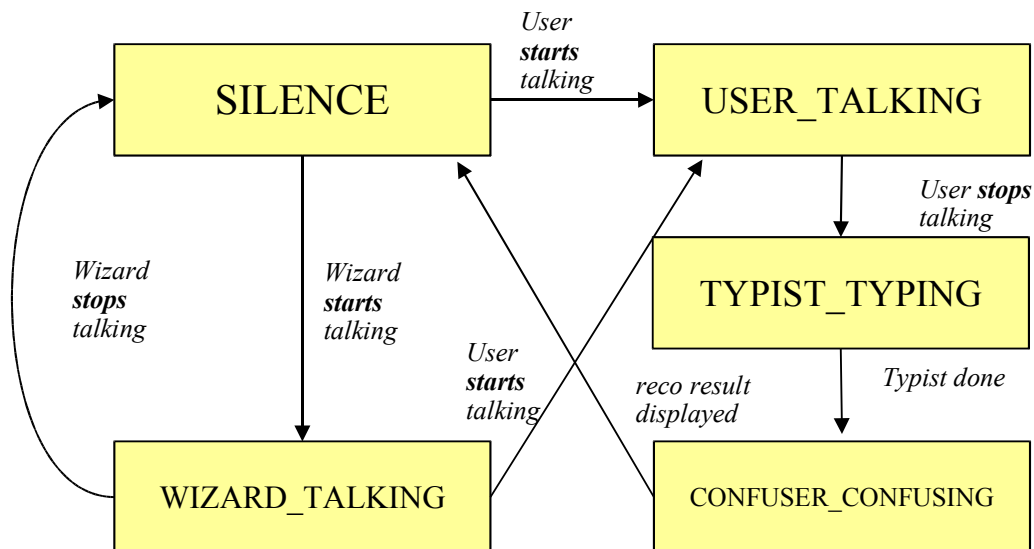


Figure 2: Transition diagram for state machine (turn-taking model)

The click map interface

Both wizard and user share the same map. They are both allowed to click at any time during the recording session on the map. When the users click on the map a circle is displayed on their map as well as on the wizards' map. Because of the time delay caused by the typist, the wizards usually do not know what was spoken, when they see the click circle. Therefore the circle is preserved on the wizards' screen until the wizards either click on a position on the map or press the clear button. A number inside the circles indicates which click was the first, the second and so on.

The user can hear the wizards speech as segmented by the endpointer. All clicks produced by the wizard are displayed simultaneously on the user's screen and vanish as soon as the wizard releases the mouse button.

Below the wizard's map are buttons to clear the screen, to display bus and tram routes and to display the locations of all hotels, restaurants or bars.

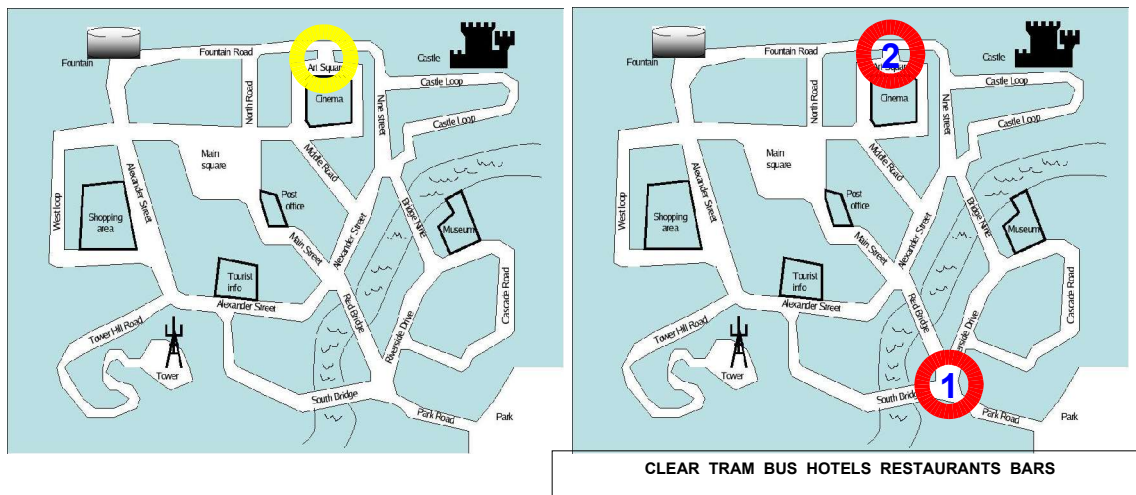


Figure 3: User's (left) and wizard's (right) maps. User clicks produced with the sentence “I am here and want to get there by public transport.”

Data collection method

This section describes the methodology used to conduct the data collection.

Tests were structured in six-hours blocks, during which one wizard interacted with six users. Each user was given 5 different tasks; thus each wizard engaged in 30 dialogues. All tasks that a wizard experienced were different, and were used in the same order from one wizard to the next. A total of 30 tasks were used.

Wizard and users were each located in different rooms, and could not see each other. Wizards and users did not interact in person before or after the tests. All tests were conducted by the same experimenter and two typists.

Wizards were recruited from a staffing agency. When wizards arrived, they were given an introduction to the task using the “Wizard Role Description”, *Form W-R* (for all forms, see Appendix A). They were then given a consent form, and asked to sign if they agreed. Finally, the same experimenter reviewed the task information given in *Form W-T*, which included a host of information about a fictitious town. An effort was made to keep the explanation consistent for all wizards.

Users were recruited from the student population of Cambridge University. When users arrived, they were given an introduction to the task using the “User Role Description”, *Form S-R*. They were then given a consent form, and asked to sign if they agreed.

Each task began with the experimenter giving the subject a task form (one of *Form S-T*), and a map (*Form S-M*). The experimenter read the task aloud, and asked if the subject had any questions. The experimenter told the user they should say “End Task” when they were finished with the task. The experimenter then left the room and began the simulation.

The experimenter stopped the experiment either when the subject said “end task”, or after a certain time limit was passed. The time limit varied depending on the task, but was generally more than 10 minutes.

Once a dialogue ended, the experimenter first obtained the wizard’s answers to the wizard’s questionnaire (*Form W-Q*).

The experimenter next interviewed the user. The experimenter first checked whether the user had filled in the appropriate boxes or made an indication on the map if required; if not, the experimenter asked the subject whether they had written down all they had learned. The map and task form were collected, and no indication was made whether the user’s submission was

correct or not. The experimenter then obtained the user's answers to the user questionnaire (*Form S-Q*). The experimenter then presented the next task (or concluded the test.)

Before each questionnaire for the user and wizard, the experimenter stressed that it was important that the subjects wrote down their honest reactions to these questions.

Corpus contents

Summary

The first 30 dialogues were recorded with bus, tram and clear buttons at the wizard interface (c,t,b). For the second lot of 89 dialogues the same interface was used and both wizard and user were urged by the experimenter to use the interactive map interface. Questions asking whether the interactive map was used and how were added to the questionnaire (c,t,b+note). For the last set of 60 dialogues three new buttons were added to the wizard's interface such that the wizard could display the locations of hotels, bars and restaurants (c,t,b,h,b,r+note). The following table gives an overview of the dialogues collected in SACTI-2:

Channel type	Wiz ID	Usr ID	No	ASR target (Task IDs)	Task instructions	Date
Sim-ASR-Click	100	101	5	None (1-1, 1-2, 1-3, 1-4, 1-5)	c,t,b	7. Jul 04
Sim-ASR-Click	100	102	5	None (2-1, 2-2, 2-3, 2-4, 2-5)	c,t,b	7. Jul 04
Sim-ASR-Click	100	103	5	None (3-1, 3-2, 3-3, 3-4, 3-5)	c,t,b	7. Jul 04
Sim-ASR-Click	100	104	5	None (4-1, 4-2, 4-3, 4-4, 4-5)	c,t,b	7. Jul 04
Sim-ASR-Click	100	105	5	None (5-1, 5-2, 5-3, 5-4, 5-5)	c,t,b	7. Jul 04
Sim-ASR-Click	100	106	5	None (6-1, 6-2, 6-3, 6-4, 6-5)	c,t,b	7. Jul 04
Sim-ASR-Click	107	108	5	Med (1-1, 1-2, 1-3, 1-4), Hi (1-5)	c,t,b + note	14. Jul 04
Sim-ASR-Click	107	109	5	Med (2-1, 2-2, 2-3, 2-4), Hi (2-5)	c,t,b + note	14. Jul 04
Sim-ASR-Click	107	110	5	Med (3-1, 3-2, 3-3, 3-4), Hi (3-5)	c,t,b + note	14. Jul 04
Sim-ASR-Click	107	111	5	Med (4-1, 4-2, 4-3, 4-4), Hi (4-5)	c,t,b + note	14. Jul 04
Sim-ASR-Click	107	112	5	Med (5-1, 5-2, 5-3, 5-4), Hi (5-5)	c,t,b + note	14. Jul 04
Sim-ASR-Click	107	113	5	Med (6-1, 6-2, 6-3, 6-4), Hi (6-5)	c,t,b + note	14. Jul 04
Sim-ASR-Click	101	114	5	Med (1-1, 1-2, 1-3, 1-4), Hi (1-5)	c,t,b + note	15. Jul 04
Sim-ASR-Click	101	115	5	Med (2-1, 2-2, 2-3, 2-4), Hi (2-5)	c,t,b + note	15. Jul 04
Sim-ASR-Click	101	116	5	Med (3-1, 3-2, 3-3, 3-4), Hi (3-5)	c,t,b + note	15. Jul 04
Sim-ASR-Click	101	117	5	Med (4-1, 4-2, 4-3, 4-4), Hi (4-5)	c,t,b + note	15. Jul 04
Sim-ASR-Click	101	118	5	Med (5-1, 5-2, 5-3, 5-4), Hi (5-5)	c,t,b + note	15. Jul 04
Sim-ASR-Click	101	119	5	Med (6-1, 6-2, 6-3, 6-4), Hi (6-5)	c,t,b + note	15. Jul 04
Sim-ASR-Click	120	121	4	Med (1-1, 1-2, 1-3, 1-4)	c,t,b + note	16. Jul 04
Sim-ASR-Click	120	122	5	Med (2-1, 2-2, 2-3, 2-4), Hi (1-5)	c,t,b + note	16. Jul 04
Sim-ASR-Click	120	123	5	Med (3-1, 3-2, 3-3, 3-4), Hi (3-5)	c,t,b + note	16. Jul 04
Sim-ASR-Click	120	124	5	Med (4-1, 4-2, 4-3, 4-4), Hi (4-5)	c,t,b + note	16. Jul 04
Sim-ASR-Click	120	125	5	Med (5-1, 5-2, 5-3, 5-4), Hi (5-5)	c,t,b + note	16. Jul 04
Sim-ASR-Click	120	126	5	Med (6-1, 6-2, 6-3, 6-4), Hi (6-5)	c,t,b + note	16. Jul 04
Sim-ASR-Click	122	127	5	Med (1-1, 1-2, 1-3, 1-4), Hi (1-5)	c,t,b,h,b,r+note	21. Jul 04
Sim-ASR-Click	122	128	5	Med (2-1, 2-2, 2-3, 2-4), Hi (2-5)	c,t,b,h,b,r+note	21. Jul 04
Sim-ASR-Click	122	129	5	Med (3-1, 3-2, 3-3, 3-4), Hi (3-5)	c,t,b,h,b,r+note	21. Jul 04
Sim-ASR-Click	122	130	5	Med (4-1, 4-2, 4-3, 4-4), Hi (4-5)	c,t,b,h,b,r+note	21. Jul 04
Sim-ASR-Click	122	131	5	Med (5-1, 5-2, 5-3, 5-4), Hi (5-5)	c,t,b,h,b,r+note	21. Jul 04
Sim-ASR-Click	122	132	5	Med (6-1, 6-2, 6-3, 6-4), Hi (6-5)	c,t,b,h,b,r+note	21. Jul 04
Sim-ASR-Click	133	134	5	Med (1-1, 1-2, 1-3, 1-4), Hi (1-5)	c,t,b,h,b,r+note	22. Jul 04
Sim-ASR-Click	133	135	5	Med (2-1, 2-2, 2-3, 2-4), Hi (2-5)	c,t,b,h,b,r+note	22. Jul 04
Sim-ASR-Click	133	136	5	Med (3-1, 3-2, 3-3, 3-4), Hi (3-5)	c,t,b,h,b,r+note	22. Jul 04
Sim-ASR-Click	133	137	5	Med (4-1, 4-2, 4-3, 4-4), Hi (4-5)	c,t,b,h,b,r+note	22. Jul 04
Sim-ASR-Click	133	138	5	Med (5-1, 5-2, 5-3, 5-4), Hi (5-5)	c,t,b,h,b,r+note	22. Jul 04
Sim-ASR-Click	133	139	5	Med (6-1, 6-2, 6-3, 6-4), Hi (6-5)	c,t,b,h,b,r+note	22. Jul 04

Table 3: Summary of corpus contents

Please note that participants 101 and 122 participated first as users and then as wizards in the experiment. Also, participants 110 and 112 participated in SACTI-1 as wizards.

The “Hi” error rate for the tasks 1-5, 2-5, 3-5, 4-5, 5-5 and 6-5 are caused by the fact that these “new” tasks are slightly different from the standard tasks on which the ASR confusion system was trained. This caused the error rate for these tasks to be higher.

File naming and organization

The file names in SACTI-2 differ slightly from those in SACTI-1. The following files are included: dialogue files containing clicks, time information and annotation, the speech signals are stored in dialogue recordings, endpointed recordings and manually segmented turn recordings.

Dialogue files are represented in the format $WWW-UUUMD-T$ (for example $002-006x1-2$), where:

1. WWW = the ID of the wizard.
2. UUU = the ID for the user.
3. M = indicates, to whom this data belongs (user, wizard or both) taking the values of u, w, x .
4. D = the first part of the ID of the task, ranging from 1 to 6 for all dialogues.
5. T = the second part of the ID of the task, ranging from 1 to 6 for all dialogues.

Each dialogue is composed of a text file and the associated audio files. The text is stored in an ANVIL file, accessible with the ANVIL tool [10]. Each dialogue includes the following files:

1. $WWW-UUUxD-T.anvil$: The ANVIL XML file, including transcriptions (described below).
2. $WWW-UUUuD-T.wav$: A 1-channel (mono) RIFF WAV audio file with the recording of the wizard's speech.
3. $WWW-UUUuD-T.wav$: A 1-channel (mono) RIFF WAV audio file with the recording of the user's speech.
4. $WWW-UUUxD-T.mov$: A 1-channel (mono) QuickTime movie with both sides of the conversation².

The WAV to QuickTime conversion was performed with [11]. For information on converting HTK WAV to RIFF WAV, see Appendix C.

Additionally there are two files containing ascii tables with summarised information about the corpus:

1. The file “dialogData.csv” summarising information about the recording sessions.
2. The file “speakers.csv” containing information related to the participants of the experiment.

² The QuickTime file is required because the ANVIL tool doesn't support simple audio file formats – only movie formats such as QuickTime. The tool used to convert from WAV to QuickTime supports only Mono (not stereo). [11]

Transcriptions & Annotation

User utterance transcription in the Simulated ASR Channel

As noted above, the users' utterances were transcribed in real time by the typist using an interface designed for this data collection. Since speed was a priority, the conventions used are intentionally narrow in scope. Further, transcriptions (especially longer transcriptions) may contain errors as speed was a conscious priority. The guidelines used by the typist are given in *Form R-T* (see Appendix A).

Wizard utterance transcription in the Simulated ASR Channel

The wizards' utterances were transcribed using PRAAT [12] off-line, after the conclusion of the dialogues. Thus the wizard utterances could be transcribed for a richer set of phenomena. This orthographic representation of the wizard's speech, corresponds to Level I transcriptions as described in [4]. We surveyed transcription conventions from Verbmobil [5], HCRC Map Task [6], and LDC [7] [8] [9]. None of the conventions met our purposes exactly. We decided on the LDC conventions in [7] because: (1) the set of tags covered the phenomena of interest, (2) LDC tools were freely available, and (3) there was existing expertise within the wider research group with the tag set.

Thus our guidelines are a subset of the LDC "Guidelines for RT-03 Transcription," dated February, 2003 [7], with two minor modifications. First, we added an indication of end-pointing errors/issues. Second, we changed the indication of self-speech to a tag which doesn't conflict with XML – i.e.: [self-speech] ... [/self-speech]. See Appendix B for the complete transcription manual.

Automatic turn annotation in the Simulated ASR Dialogues

We were interested to know, at each point in the dialogues, who has the "floor" of the conversation. As an attempt to express this, we have included a track in the dialogues which shows this. This track was added by an automatic procedure (i.e., it was not hand-annotated), according to rules described below.

In the simulated ASR channel dialogues, we defined a turn as the maximal sequence of non-silent utterances from the same participant during which the other participant doesn't speak. Because only one participant can speak at one time, this decision procedure is easy to define. The end result is a sequence of turns which is guaranteed to alternate between the user and wizard, containing at least one non-silent utterance for each turn, and possibly silence between and within each turn.

Task-completion annotation

Included with each dialogue is an objective assessment of task completion. Task completion was determined after the end of the session by examining the user's submitted map and task form – i.e., the content of the dialogues themselves were not considered. The same individual performed all task completion assessments. Task completion has been graded with respect to Precision and Recall.

- Task Recall: Indicates the amount of information collected with respect to the task *without* assessing its accuracy.
 - ◆ 3: All information gathered
 - ◆ 2: Minor pieces of information missing
 - ◆ 1: Major pieces of information missing
 - ◆ 0: Nothing attempted

- **Task Precision:** Indicates whether the information the user submitted on their map and task is *accurate*, without assessing what portion of the task it satisfied. Several guidelines that were followed are included in each category.
 - ◆ **3:** All information provided is correct. For example, establishment locations like hotels and bars must be on the correct "block", but may be on the opposite side of the street.
 - ◆ **2:** Most information is correct. For example, establishment locations may be on the "next" block.
 - ◆ **1:** Major inaccuracy. For example, any route description (such as a bus or tram) which goes via an incorrect street
 - ◆ **na:** Nothing attempted (i.e., Task Recall = 0)

Dialogue details

Each dialogue is stored as an ANVIL-format XML file. The simulated ASR dialogues are defined in the XML file "click-anvil-spec.xml". Each file (dialogue) contains 8 ANVIL "tracks" and 1 ANVIL "set".

1. states sequence as obtained directly from logfile,
2. turns sequence,
3. clicks of the user,
4. clicks of the wizard,
5. button pressed by the wizard
6. endpoint user turns
7. endpoint wizard turns
8. manual segmentation and transcription of wizard's speech

Resolution for time stamps of different tracks

The recordings have been done on three different machines with separate clocks, that were synchronised by the network. Therefore the time resolution and the synchronisation errors are different for each pair of tracks.

A **time resolution less than 1ms** was obtained for data that was recorded on the same machine such as

- endpoint user utterances and user clicks
- endpoint and manually segmented wizard utterances, wizard clicks and wizard button events

A **time difference less than 50 ms** was achieved with speech recorded on different machines. These are track pairs such as:

- wizard's speech vs. user's speech
- wizard's clicks vs. user's clicks
- This time difference applies also to the offset between the speech signals of wizard and user, which were mixed together to produce the movie.

A **time difference less than 500ms** applies when two tracks were recorded on different machines and commands take up time while travelling through the network. This is the case for the time difference between the states track and all other tracks, except the turns track, which is directly derived from the states track. This implies that these differences apply for the time stamps of the state machine and all the speech signals.

“states” track

In the simulated ASR dialogues, one ANVIL “track” called “states” is used to show the progression of the state machine. Different colours represent the different states. Wizard transcriptions are in the transcription attribute of WIZARD_TALKING states. User transcriptions are in the transcription attribute of TYPIST_TYPING states. The result of the ASR confusion process for the users’ utterances (i.e., the result displayed on the wizard’s screen) is given as the confused attribute of CONFUSER_CONFUSING states. The time resolution of the States track vs. the speech signal and all other tracks is less than 500ms. States have the following properties:

Property	Description	Example(s)
type	Indicates the state taken directly from the statemachine.	SILENCE, USER_TALKING, WIZARD_TALKING, TYPIST_TYPING, CONFUSER_CONFUSING, NULL
wavFile	For USER_TALKING and WIZARD_TALKING states, indicates the corresponding wav file.	wiz-000--2004-02-06-11-20-15.wav
transcription	For TYPIST_TYPING and WIZARD_TALKING states, indicates the actual words spoken.	NOT TOO EXPENSIVE
confused	For CONFUSER_CONFUSING states, indicates the text placed on the wizard's screen, after the simulated ASR confusion.	NIGHT TOO EXPENSIVE
epErrorStart	For WIZARD_TALKING states, indicates if the transcriptionist observed that the wizard was cut off at the end of their utterance (may indicate the user began talking over the wizard, or an end-pointer problem).	True, False
epErrorEnd	For WIZARD_TALKING states, indicates if the wizard was cut off at the beginning of their utterance (most likely due to end-pointer problem).	True, False
transComments	For WIZARD_TALKING states, indicates any comments added by the transcriptionist.	he coughs at the end
cmd	Indicates the statemachine low-level "command" which caused the transition <i>out of</i> this state and into the next state.	start, stop
id	Indicates the identity component which generated <i>cmd</i> , causing the transition <i>out of</i> this state and into the next state.	userears, wizardears
errorMsg	Indicates contents (if any) of the error tag in the statemachine logs. This indicates a low-level issue with the statemachine, usually events arriving out of order due to network traffic. This does not indicate that the participants experienced a problem with the simulation.	No handler in transition hash for state=USER_TALKING, id=tm.typist, cmd=silence, msg=silence
next	Indicates the <i>type</i> of the next state.	Same as <i>type</i> , above.

Table 4: Properties of the *states* ANVIL track in the Simulated-ASR dialogues

“turns” track

The “turns” track expresses the “turn” information described above for the simulated ASR channel dialogues. A “turn” in the simulated ASR channel dialogues is defined as an ANVIL “span” of elements in the “states” track. The ANVIL track “turns” has the following property:

Property	Description	Example(s)
who	Identity of the active participant in this turn.	Wizard, User

Table 5: Properties of the *turns* ANVIL track in the Simulated-ASR dialogues

“userUttsEP”, “wizardUttsEP” and “wizardUtts” track

The Wizard’s and User’s utterances are shown in three separate tracks. The ANVIL tracks “userUttsEP” and “wizardUttsEP” show the turns as segmented by the automatic endpointer with the respective transcription. In these tracks the user transcriptions were spell-checked and corrected (as opposed to the states tier, where they were included as typed by the typist). Since the endpointer does not work perfect it is very likely that words were cut or parts of the utterance were left out now and then.

The ANVIL track “wizardUtts” presents a manual segmentation and transcription, that was done off-line after the recordings. This transcription is more accurate than the endpointer transcriptions. The track “wizardUttsEP” was created automatically from this tier using forced Viterbi-alignment. All “utts” tracks are time aligned with their respective speech signal within 1ms. The time difference to the states track is less than 500ms.

All three tracks are instances of the ANVIL Group “utts” and have the following properties:

Property	Description	Example(s)
transcription	Indicates the actual words spoken, as transcribed by the transcriptionist or typist (spell-checked)	userUttsEP: NOT TOO EXPENSIVE wizardUttsEP: NOT TOO EXPENSIVE wizardUtts: not too expensive
wavFile	wavFile containing the speech segment corresponding to the transcription	user-087—2004-07-15-11-48-18.wav, wiz-072—2004-07-15-11-48-15.wav, DialogueWizRec2004-07-15--11-03-54_006_02.wav
utterance	This attribute does not provide any information; it is simply used to set the colour of the track for visual clarity.	User, Wizard

Table 6: Properties of the *utts* ANVIL track

“userClicks” track and “wizardClicks” track

The wizard’s and user’s clicks are shown on separate tracks. The time stamps indicate when the mouse button was pressed and released. The attributes indicate the coordinates on the map. The time stamps are accurate within 1ms to the corresponding speech signal. Both tracks “userClicks” and “wizardClicks” are instances of the ANVIL Group “clicks” and have the following properties:

Property	Description	Example(s)
xpress	Indicates the x-position of the pointer on the map when the mouse button was pressed	1, 2, 3, ...
ypress	Indicates the y-position of the pointer on the map when the mouse button was pressed	1, 2, 3, ...
xrelease	Indicates the x-position of the pointer on the map when the mouse button was released	1, 2, 3, ...
yrelease	Indicates the y-position of the pointer on the map when the mouse button was released	1, 2, 3, ...

Table 7: Properties of the *userClicks* and *wizardClicks* ANVIL track

“wizardButtons” track

The track called “wizardButtons” shows when the wizard pressed a button on his graphical user interface, to display bus or tram lines or to show the positions of hotels, bars or restaurants. The button events have no duration (start time = end time). The time stamps are accurate within 1ms to the corresponding speech signal. The track “wizardButtons” is an instance of the ANVIL Group “buttons” and has the following properties:

Property	Description	Example(s)
utterance	This attribute does not provide any information; it is simply used to set the colour of the track for visual clarity.	Wizard, User
button	Indicates which button was pressed	bus, tram, clear, hotel, bar, restaurant

Table 8: Properties of the *wizardButtons* ANVIL track

Dialogue Properties set

In addition to transcriptions, each dialogue file (for both the simulated ASR and full-audio dialogues) contains a host of raw data associated with that dialogue. This data is stored as an ANVIL “set” called “dialogProps”, which includes:

Property	Description	Example(s)
FileName	The file name (minus the extension like .anvil or .mov) of this file.	001-001x1-1
SoundfileUser	The file name (minus the extension .wav) of the recording of the user	001-001u1-1
SoundfileWiz	The file name (minus the extension .wav) of the recording of the wizard	001-001w1-1
Date	The date on which this dialogue was conducted.	06-Feb
WizardID	The ID of the wizard in this corpus. These are IDs that identify people. Participants 101 and 122 participated first as users and then as wizards in the experiment. Participants 110 and 112 participated in SACTI-1 as wizards.	100, 122, ...
UserID	The ID of the user in this dialogue. (Participants 110 and 112 were users in SACTI-2 and participated in SACTI-1 as wizards.)	103, 104, ...
Native	Indicates whether the user is a native or non-native speaker of English. Here we define Native Speaker to mean any one who lists any dialect of English as their mother tongue.	Native, NonNative
ErrorRate	The ASR error rate target for this dialogue. For simulated ASR channels, <i>Low</i> , <i>Med</i> , and <i>Hi</i> indicate the target level of corruption; <i>None</i> indicates that the transcription is passed directly to the wizard unaltered.	None, Low, Med, Hi
TaskID	The ID of the task used in this dialogue. There are a total of 30 tasks.	1, 2, ... 30
UserOrder	The order of this dialogue as experienced by the user. Each user undertook 5 dialogues.	1, 2, 3, 4, 5
WizardOrder	The order of this dialogue as experienced by the wizard. Each wizard undertook 30 dialogues with 6 different users (5 dialogues per user).	1, 2, ... 6
WizardUser	The order of this user as experienced by this wizard. Each wizard interacted with 6 different users.	1, 2, 3, 4, 5, 6
TaskPrecision	A grading of precision of the answer provided by the user. See discussion of task completion, above.	na, 1, 2, 3
TaskRecall	A grading of the recall of the answer provided by the user. See discussion of task completion, above.	0, 1, 2, 3
Experimenter-Terminated	True indicates whether the experimenter terminated the experiment. False indicates that the user ended the experiment (by saying something like "end task").	True, False
WizardLikert {1..9}	Wizard's response to Likert-scored question <i>n</i> on Form W-Q for this dialogue.	1, 2, ... 7
CommWizard Likert6	free text as written by the wizard	“show bus/tram routes and locations”
UserLikert {1..8}	User's response to Likert-scored question <i>n</i> on Form S-Q for this dialogue.	1, 2, ... 7
CommUserLikert6	free text entered by the user	“didn't use”

Table 9: Contents of *dialogProps* ANVIL set included with each dialogue

Example Simulated ASR Channel ANVIL file

Following is an excerpt of an ANVIL file.

The ANVIL file first references the specification file, `click-anvil-spec.xml`. The ANVIL file next references the associated audio (video) file `133-137x4-2.mov`.

In this example, the wizard says nothing, then “hello”. Then the user says “I NEED PUBLIC TRANSPORTATION FROM THE TOURIST OFFICE TO MY HOSTEL”, which is displayed on the wizard’s screen as “UH ME AWAY TRANSPORTATION FIND A TOURIST OFFICE TO MY HOSTEL.” ...

```
<?xml version="1.0" encoding="ISO-8859-1"?>
<annotation>
  <head>
    <specification src="click-anvil-spec.xml" />
    <video src="133-137x4-2.mov" />
  </head>
  <body>
    <track name="states" type="primary">
      <el index="0" start="0.0000" end="0.7267">
        ...
      </el>
      <el index="1" start="0.7267" end="2.2873">
        <attribute name="wavFile">wiz-506--2004-07-22-15-16-15.wav</attribute>
        <attribute name="next">SILENCE</attribute>
        <attribute name="epErrorStart"></attribute>
        <attribute name="cmd">stop</attribute>
        <attribute name="transComments"></attribute>
        <attribute name="transcription">HELLO</attribute>
        <attribute name="type">WIZARD_TALKING</attribute>
        <attribute name="id">ep.userears</attribute>
        <attribute name="epErrorEnd"></attribute>
      </el>
      ...
      <el index="3" start="4.1273" end="8.7523">
        <attribute name="wavFile">user-334--2004-07-22-15-16-18.wav</attribute>
        <attribute name="next">TYPIST_TYPING</attribute>
        <attribute name="cmd">stop</attribute>
        <attribute name="type">USER_TALKING</attribute>
        <attribute name="id">ep.typistears</attribute>
      </el>
      <el index="4" start="8.7523" end="17.9449">
        <attribute name="next">CONFUSER_CONFUSING</attribute>
        <attribute name="cmd">transcription</attribute>
        <attribute name="transcription">
          I NEED PUBLIC TRANSPORTATION FROM THE TOURIST OFFICE TO MY HOSTEL
        </attribute>
        <attribute name="type">TYPIST_TYPING</attribute>
        <attribute name="id">tm.typist</attribute>
      </el>
      <el index="5" start="17.9449" end="18.6471">
        <attribute name="next">SILENCE</attribute>
        <attribute name="confused">
          UH ME AWAY TRANSPORTATION FIND A TOURIST OFFICE TO MY HOSTEL
        </attribute>
        <attribute name="cmd">confused</attribute>
        <attribute name="type">CONFUSER_CONFUSING</attribute>
        <attribute name="id">confuser</attribute>
      </el>
      ...
    </track>
    <track name="turns" type="span" ref="states">
      <el index="0" start="1" end="1">
        <attribute name="who">Wizard</attribute>
      </el>
      <el index="1" start="3" end="5">
        <attribute name="who">User</attribute>
      </el>
    </track>
  </body>
</annotation>
```

```

...
</track>
<track name="clicks.userClicks" type="primary">
</track>
<track name="clicks.wizardClicks" type="primary">
  <el index="0" start="46.0562999248505" end="51.5840499401093">
    <attribute name="xpress">360</attribute>
    <attribute name="ypress">177</attribute>
    <attribute name="xrelease">360</attribute>
    <attribute name="yrelease">177</attribute>
    <attribute name="utterance">Wizard</attribute>
  </el>
  ...
</track>
<track name="buttons.wizardButtons" type="primary">
  <el index="0" start="31.5882999897003" end="31.5882999897003">
    <attribute name="button">bus</attribute>
    <attribute name="utterance">Wizard</attribute>
  </el>
  ...
</track>
<track name="utts.userUttsEP" type="primary">
  <el index="0" start="5.12674981880188" end="9.95174981880188">
    <attribute name="utterance">User</attribute>
    <attribute name="transcription">
      I NEED PUBLIC TRANSPORTATION FROM THE TOURIST OFFICE TO MY HOSTEL
    </attribute>
    <attribute name="wavFile">user-334--2004-07-22-15-16-18.wav</attribute>
  </el>
  ...
</track>
<track name="utts.wizardUttsEP" type="primary">
  <el index="0" start="1.73674995231628" end="3.49924995231628">
    <attribute name="utterance">Wizard</attribute>
    <attribute name="transcription">HELLO</attribute>
    <attribute name="wavFile">wiz-506--2004-07-22-15-16-15.wav</attribute>
  </el>
  ...
</track>
<track name="utts.wizardUtts" type="primary">
  <el index="0" start="1.50713526109482" end="2.537125">
    <attribute name="utterance">Wizard</attribute>
    <attribute name="transcription">hello</attribute>
    <attribute name="wavFile">
      DialogueWizRec2004-07-22-10-10-49_018_00.wav
    </attribute>
  </el>
  ...
</track>
<set name="dialogProps">
  <el index="0">
    <attribute name="FileName">133-137x4-2</attribute>
    <attribute name="SoundfileUser">133-137u4-2</attribute>
    <attribute name="SoundfileWiz">133-137w4-2</attribute>
    <attribute name="Date">22-July</attribute>
    <attribute name="WizardID">133</attribute>
    <attribute name="UserID">137</attribute>
    <attribute name="ErrorRate">Med</attribute>
    <attribute name="TaskID">4-2</attribute>
    <attribute name="UserOrder">2</attribute>
    <attribute name="WizardOrder">17</attribute>
    <attribute name="WizardUser">4</attribute>
    <attribute name="TaskPrecision">3</attribute>
    <attribute name="TaskRecall">2</attribute>
    <attribute name="ExperimentorTerminated">FALSE</attribute>
    <attribute name="Native">Native</attribute>
    <attribute name="WizardLikert1">6</attribute>
    ...
    <attribute name="WizardLikert6">6</attribute>
    <attribute name="CommWizardLikert6">showing bus / tram routes + hotel
location</attribute>

```

```

<attribute name="WizardLikert7">4</attribute>
<attribute name="WizardLikert8">No</attribute>
<attribute name="WizardLikert9">5</attribute>
<attribute name="UserLikert1">7</attribute>
...
<attribute name="UserLikert6">2</attribute>
<attribute name="CommUserLikert6">showing all routes</attribute>
<attribute name="UserLikert7">2</attribute>
<attribute name="UserLikert8">6</attribute>
</el>
</set>
</body>
</annotation>

```

References

- [1] Matthew Stuttle, Jason D. Williams, and Steve Young. (2004). "A Framework for Dialogue Data Collection with a Simulated ASR Channel." Submitted for publication.
- [2] Jason D. Williams and Steve Young. (2004). "Characterizing Task-Oriented Dialog using a Simulated ASR Channel." Submitted for publication.
- [3] Gabriel Skantze. (2003). "Exploring human error handling strategies: implications for spoken dialogue systems." *Proc. Error Handling in Spoken Dialogue Systems, ISCA Tutorial and Research Workshop*, Château d'Oex, Vaud, Switzerland, August 28-31, 2003, pp 71-76.
- [4] J. P. French. (1992) "Transcription proposals: multi-level system." Working Paper, University of Birmingham, October 1992. NERC-WP 4-50.
- [5] Verbmobil transcription conventions. http://www.is.cs.cmu.edu/trl_conventions/
- [6] HCRC Map Task transcription conventions.
http://www ldc.upenn.edu/Catalog/docs/hcrc_map/editorl.sgm
- [7] LDC Transcription conventions for Broadcast News, RT-03.
http://ldc.upenn.edu/Projects/Transcription/rt-03/RT_Transcription_V2.2.pdf
- [8] LDC Transcription conventions for Switchboard task.
http://www.isip.msstate.edu/projects/switchboard/doc/transcription_guidelines/transcription_guidelines.pdf
- [9] LDC Transcription conventions for HUB5 and SPINE.
<http://www ldc.upenn.edu/Projects/SPINE/transconv.html>
- [10] M. Kipp. (2001). ANVIL – A Generic Annotation Tool for Multimodal Dialogue. *Proc. Eurospeech*
- [11] Malcolm Slanley. (1999). Interval Technical Report #1999-066. Interval Research.
<http://web.interval.com/papers/1999-066/>
- [12] Boersma, Paul and Weenink, David (1992–2001). Praat: A system for doing phonetics by computer. Available from www.praat.org.
- [13] Williams, Jason D. and Young, (2004) The *SACTI-1* Corpus: Guide for Research Users University of Cambridge, Department of Engineering, Technical Report: CUED/F-INFENG/TR482

Appendix A: Experimentation documents

The experimental documents are given a name consisting of two letters, like *X-Y*. The first letter indicates who the document is intended for:

- *S* for Subject-facing forms
- *W* for Wizard-facing forms
- *T* for Typist-facing forms
-

The second letter indicates the contents of the document:

- *T* for Task information
- *M* for Map
- *R* for Role Description
- *Q* for Questionnaire

Finally, if the document name ends in *a*, this indicates the document was altered to encourage the user to use the interactive map interface.

Experiments	Who	Name	Description
All dialogues	Subject	<i>Form S-T</i>	Tasks
	User	<i>Form S-M</i>	Subject Map
	Wizard	<i>Form W-T</i>	Task information
		<i>Form W-M</i>	Wizard Map
	Typist	<i>Form T-R</i>	Role description typist
1 st set of 30 dialogues	Subject	<i>Form S-R</i>	Role description subject
	User	<i>Form S-Q</i>	Questionnaire subject
	Wizard	<i>Form W-R</i>	Role description wizard
		<i>Form W-Q</i>	Questionnaire wizard
2 nd set of 89 dialogues	Subject	<i>Form S-R</i>	Role description subject
	User	<i>Form S-Q-a</i>	Questionnaire subject asking for clicks
	Wizard	<i>Form W-R-a</i>	Role description wizard urging to use clicks
		<i>Form W-Q-a</i>	Questionnaire wizard asking for clicks
3 rd set of 60 dialogues	Subject	<i>Form S-R</i>	Role description subject
	User	<i>Form S-Q-a</i>	Questionnaire subject asking for clicks
	Wizard	<i>Form W-R-a</i>	Role description wizard urging to use clicks
		<i>Form W-Q-a</i>	Questionnaire wizard asking for clicks

Table 10: Summary of experimental documents used in corpus collection

TASK 1-1

Form S-T

Finding a restaurant close to the Royal hotel

You're staying at the Royal hotel, and need to find a restaurant for dinner. You'd prefer something not too expensive.

Please submit a map showing the location of the restaurant you select, and fill in the boxes below.

Name of restaurant
Average price per person
Type of food

TASK 1-2

Form S-T

Going out for Pizza

You have a craving for a pizza tonight. Find out if there is a suitable restaurant, and if so, get their phone number so you can make a reservation.

Please submit a map showing the location of the restaurant, and fill in the boxes below.

Name of restaurant

Phone number

TASK 1-3

Form S-T

Planning out a day

You're trying to plan out your activities for the day. Please find the opening hours of the following tourist attractions.

Please fill in the boxes below.

Site	Opening hours
Museum	
Castle	
Tower	

TASK 1-4

Form S-T

Getting across town

You're currently at the castle, and would like to get the tower as quickly as possible using public transportation.

Please show the route on a map. Also, please fill in the boxes below. (Use as many rows as you need – you may not need all of the rows).

Leg of journey	Form of public transportation used for this leg of the journey (e.g., bus)	Cost
1		
2		
3		
4		
TOTAL COST		

TASK 1-5

Form S-T

Driving around

You are approaching Talk Town in your car. You have booked a room in Hotel Primus for the night, but you are already late for a 2 hour meeting in a restaurant in West Loop.

What is the name of the restaurant?

Find the quickest way to the restaurant, but keep in mind that you must park your car.

Please submit a map with the route and the location of where you intend to park your car and fill in the box below.

Name of restaurant

TASK 2-1

Form S-T

Finding an upmarket bar

You are entertaining an important client tonight. Find a suitable, up-market bar.

Please show the location of your choice on the map, and please fill in the boxes below.

Name of bar

Brief description & price range

TASK 2-2

Form S-T

Finding the cost of tourist attractions

You're considering your options for your visit, and trying to work out a budget.

Find out how much the following sites cost. Please fill in the boxes below.

Site	Cost per person
Castle	
Tower	
Fountain	
Museum	

TASK 2-3

Form S-T

Finding the perfect hotel

You're looking for a hotel for you and your travelling partner that meets a number of requirements.

You'd like the following:

1. En suite rooms
2. Quiet rooms
3. As close to the main square as possible

Given those desires, find the least expensive hotel. You'd prefer not compromise on your requirements, but of course you will if you must!

Please indicate the location of the hotel on the map and fill in the boxes below.

Name of accommodation

Cost per night for 2 people

TASK 2-4

Form S-T

Running errands

You're meeting a friend for dinner at the Hotel President, which is on Castle loop, near Bridge Nine.

After a very successful day of shopping at the shopping centre, you've bought so many items you need to take public transportation to get to the hotel. However, on the way you need to stop & post a letter at the post office.

Find out what public transportation will take you from the shopping centre to the post office, and then from the post office to the hotel.

Please submit a map showing your route.

TASK 2-5

Form S-T

Using your car in the town

You have been shopping in the shopping centre, where your car is parked at the moment. You want to pick up a heavy parcel from the post office and you want to meet a friend at the Royal Hotel and have dinner with him. After that you will drive home to your village.

After you have made a plan, please submit a map showing the location of the Royal Hotel, the restaurant and the places, where you will park your car. Please fill in the boxes below.

Activity	Location of car park
shopping at shopping centre	shopping centre

TASK 3-1

Form S-T

Finding “Café Blu”

There is a bar in town called “Café Blu”. You’re meeting a friend there but are running late.

Find out where the Café is located, and get their phone number.

Please submit a map showing its location, and fill in the box below.

Phone number of Café Blu

TASK 3-2

Form S-T

Finding out about public transportation

You're trying to understand how the public transportation works.

On your map, trace out the public transportation routes, labelling the route numbers.

When finished, please submit the map showing the routes.

TASK 3-3

Form S-T

Findings a children’s film & posting a letter

You’re babysitting a young child, and have just finished lunch. It’s now 1:15 PM.

This afternoon, you’d like to take the child to see a film suitable for young children, and you also need to post a letter. The queue at the post office is often long, so you’ll need at least 20 minutes at the post office.

You’re now at the cinema. With the child, it takes about 20 minutes to walk from the cinema to the post office, or visa-versa.

Make a schedule for your afternoon. Please fill in the boxes below.

Time	Activity

Name of film

TASK 3-5

Form S-T

Posting a heavy parcel and watching a film

It is about noon and you are approaching Talk Town in your car. You want to post a heavy parcel at the post office. After that you have plenty of time before you meet a friend to go to the cinema in the evening. After the movie you will drive back home to your village.

After you have made a plan for your visit of the town, please submit a map showing your activities and the places, where you will park your car. Please fill in the boxes below.

Title and showing time of the movie

Activity	Location of car park

TASK 4-1

Form S-T

Finding budget accommodation

You're travelling on a budget. You've just arrived in town, and are looking for the least expensive accommodation.

Please submit a map showing the name of the accommodation you select, and how much it will cost per night for 1 person.

Name of accommodation

Price per night for 1 person

TASK 4-2

Form S-T

Getting to your hostel

You've just left the tourist information booth, and they have made a reservation for you at the Youth Hostel.

Find out what public transportation can get you from the tourist office to your hostel, including how long it will take, and much it will cost.

Please submit a map showing the location of your hostel, and also tracing out the route of the public transport. Please also fill in the boxes below:

How much will the public transportation cost?

How long will it take on the public transportation to get from the tourist information booth to your hostel?

TASK 4-3

Form S-T

What's on at the cinema

You'd like to see a film tonight. Get a list of what's on at the cinema.

Please fill in the boxes below.

Film name	Time(s)	Rating	Ticket price per person

TASK 4-4

Form S-T

Planning out a day out

You're staying at the Youth Hostel. You'd like to take a tour of the castle and tower, seeing as much of both as possible, but definitely seeing some of both.

In an effort to see more of the town, you're getting around today strictly by walking.

You're starting from your hostel, and it's now Noon.

Walking around town takes the following amounts of time:

Walking between...	Time
Hostel and castle	20 minutes
Hostel and tower	20 minutes
Castle and tower	40 minutes

Find out when the tours operate, and plan out your activities for the day.

Please fill in the boxes below, including when you'll be walking between various sites.

Time	Activity
Noon	Starting from the Youth Hostel, walk to...

TASK 4-5

Form S-T

Two people visiting the town

You are two people in a car approaching Talk Town. You want to visit the museum while the other person wants to go shopping. After two or three hours you will meet again, have some food and drive back home to your village.

After you have made a plan for the afternoon and the evening, please submit a map showing your activities and the places, where you will park your car. Please fill in the boxes below.

Name of the restaurant

Activity	Location of car park

TASK 5-1

Form S-T

Going for a curry

You've just arrived in town, and you have a craving for a curry tonight. Find out if there is a suitable restaurant, and if so, get their phone number so you can make a reservation.

Please submit a map showing the location of the restaurant, and fill in the boxes below.

Name of restaurant

Phone number

TASK 5-2

Form S-T

Finding a hotel

You've just arrived in town and are looking for a moderately priced hotel (double room).

Please submit two items:

6. A list of hotels and their price ranges
7. A map showing the location of the hotels.

Hotel name	Price range

TASK 5-3

Form S-T

Planning out a day

You're trying to plan out your activities for the day. Please find the opening hours, and cost, of the following tourist attractions.

Please fill in the boxes below.

Site	Cost per person	Opening hours
Tower		
Museum		
Fountain		

TASK 5-5

Form S-T

Meeting at the tower

You are approaching Talk Town in your car. You want to meet a friend at the iron tower. You plan to have a drink in a pub before you go for dinner in a restaurant. After dinner you will drop off your friend at his hotel in castle loop, before you drive back home to your village.

After you have made a plan for the afternoon and the evening, please submit a map showing your activities and the places, where you will park your car. Please fill in the boxes below.

Name of the bar

Name of the restaurant

Name your friends hotel

Activity	Location of car park

TASK 6-1

Form S-T

Finding a bar or café close to the fountain

You're at the fountain, and would like to find the nearest bar or cafe.

Please submit a map showing the location of the bar or café you select, and fill in the boxes below:

Name of bar or cafe

Brief description of bar or cafe

TASK 6-2

Form S-T

Getting back to your hotel

You're staying at the Hotel Primus, which is located at the corner of Alexander street and West loop.

At the moment, you're in the park, and need to get back to your hotel.

It is now 4:30 PM.

Find out what public transportation options exist, including their schedule and how much it will cost.

Please fill in the boxes below.

When & where will you get on the public transportation?

When & where will you get off the public transportation?

TASK 6-3

Form S-T

Restaurant locations

You are evaluating your options for a restaurant this evening, and would like to get a list of the restaurants in the town.

Please submit two items:

1. A map showing the locations of each restaurant
2. The names and average prices of the restaurants

Number on map	Restaurant name	Average price

TASK 6-4

Form S-T

Planning a bar crawl

You are planning a bar crawl for a group of friends travelling together.

You are staying at Hotel Primus, and will start and end here.

You'll be starting at about 8 AM, and finishing at about 1 AM (hopefully!)

You'd like to visit about 4 bars.

You'll be walking, so try to minimize the distances between the bars.

Ideally, you'd prefer inexpensive bars.

Once you've formulated your plan, please submit two items:

1. A map showing the route you'll take, and
2. A list of the bars you'll visit

Bar name

TASK 6-5

Form S-T

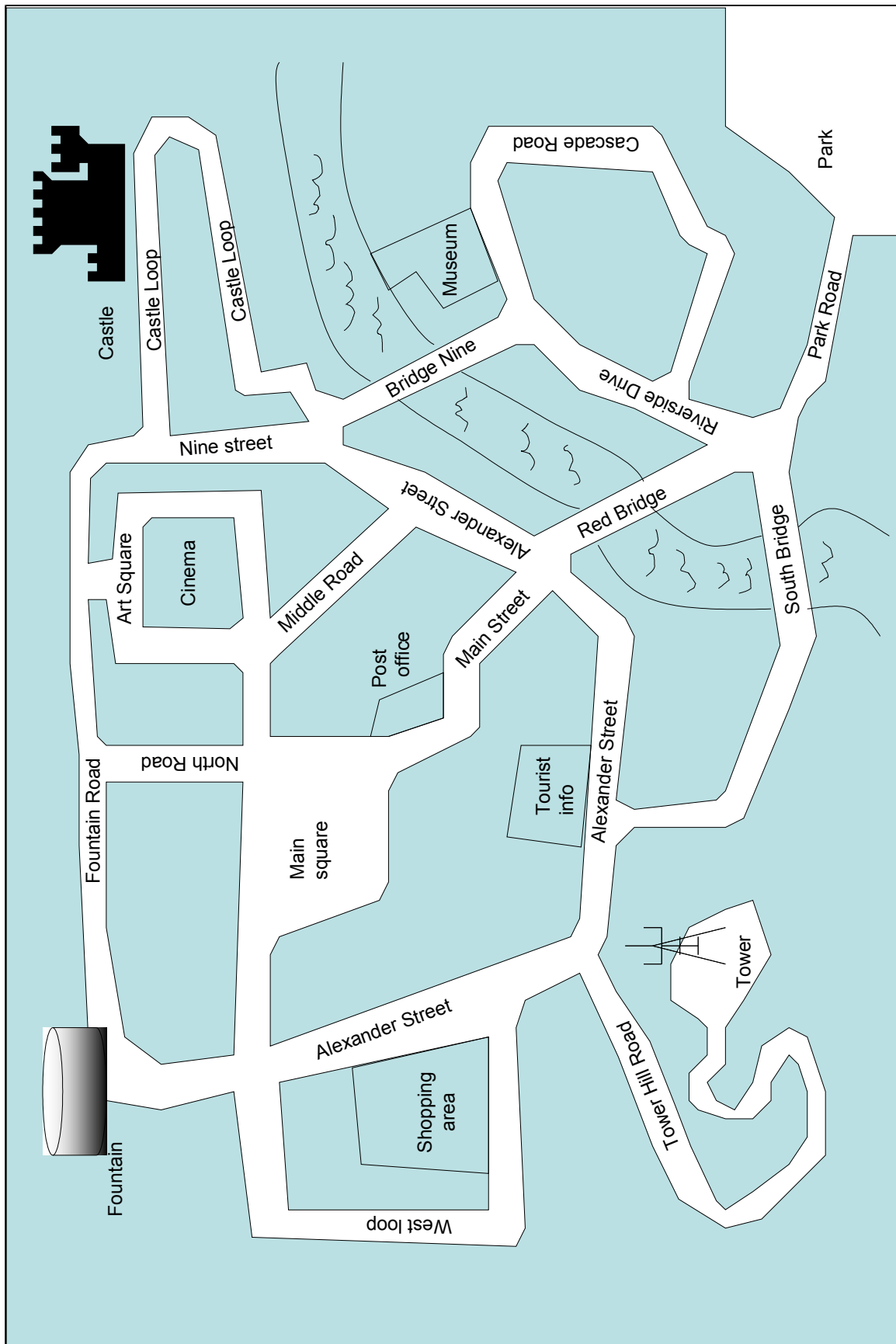
Picking people up

You agreed to be the driver this evening. You are picking up 3 people from different places and will drive them home to your village. You just had dinner at a restaurant, called "The Grand". Your car is in the car park below the museum. One of your friends is waiting in the "Bar Metropol", another in a Pub called Murphies, another in front of the Cinema.

Please draw the route you want to drive to pick up your friends (please mark pick up places) and leave the town for your village. Keep in mind that one-way streets and areas that are closed for car traffic are not marked in your map. Please write them down in the boxes below.

areas closed for car traffic

one-way streets



TOWN INFORMATION

Form W-T

Hotel listing

(H1) HOTEL PRIMUS

Address: Alexander Street & West Loop

Tel.: (2)094-227

Single Room: GBP 20 / night

Double Room: GBP 30 / night

Big, bright, clean rooms, some en suite, but with some noise from the busy street below

(H2) HOTEL PRESIDENT

Address: Castle Loop

Tel.: (7)192-277

Single Room: GBP 35 / night

Double Room: GBP 45 / night

Note: two night minimum stay

Business-style hotel, efficiently staffed. All rooms en suite. Breakfast included.

(H3) ROYAL HOTEL

Address: Cascade Road

Tel.: (7)027-003

Single Room: GBP 50 / night

Double Room: GBP 80 / night

Renovated nineteenth-century palace, featuring private balconies. Spacious well-appointed rooms, many of which offer stunning views. All rooms en suite, and breakfast is included.

(H4) YOUTH HOSTEL

Address: Main Square

Tel.: (7)281-446

Per bed: GBP 8 / person

Hostel serving students and budget travellers. Bunk-beds in shared rooms with shared toilet facilities.

(H5) ART HOUSE HOTEL

Address: Fountain Road

Tel.: (7)252-996

Single Room: GBP 15 / night

Double Room: GBP 20 / night

Basic hotel, frequented by artists and musicians. Rooms are decorated in bright colours, but few offer views, and all toilet facilities are shared. Breakfast not included.

(H6) ALEXANDER HOTEL

Address: Alexander Street

Tel.: (7)874-874

Single Room: GBP 18 / night

Double Room: GBP 22 / night

Very comfortable, pleasantly decorated (but small) rooms, with a very friendly staff. Shared toilet facilities. Breakfast included.

Restaurant listing

(R1) *BOCHKA*

Address: Main square

Tel.: (2)095-252

Price: GBP 8 / person

A tavern open around the clock, serving excellent sandwiches, soups, tavern meals, and salads.

(R2) *SIBERIAN TIGER*

Address: West loop

Tel.: (7)095-926

Price: GBP 38 / person

Savoury Russian cooking, accompanied by first-class music by Bolchoi musicians.

(R3) *SAINT PETERSBURG*

Address: Park Road

Tel.: (7)812-311

Price: GBP 20 / person

The most interesting restaurant in town serving indian fusion food chic decoration.

(R4) *NOBLE NEST*

Address: Fountain Street & North Road

Tel.: (7)812-312

Price: GBP 14 / person

Relaxed, family-style restaurant serving Chinese food

(R5) *THE GRAND*

Address: Cascade Road

Tel.: (7)812-329

Price: GBP 50 / person

One of the only two "Leading Restaurants" in the country. Impeccable service. Known for their caviar.

(R6) *CHEZ SERGU*

Address: Middle Road and Art Square

Tel.: (7)812-465

Price: GBP 29 / person

Classic French restaurant, with extensive wine list.

Bars listing

(B1) BAR METROPOL

Address: Main square

Tel.: (7)812-310

Price: Moderate

Upscale, modern décor with live, soft jazz

(B2) CAFÉ BLU

Address: Alexander Street

Tel.: (7)812-314

Price: Inexpensive

Loud music, cheap beer, dancing – open until 5 AM

(B3) BUFFALO BILLS

Address: Alexander Street

Tel.: (7) 812-329

Price: Inexpensive

Known for margaritas, tequilas, and Mexican beers

(B4) EUROPA

Address: South Bridge

Tel: (2) 512 3283

Price: Expensive

Tiny bar with exclusive clientele. Being on the guest list is a must.

(B5) PLANTER'S

Address: South Bridge

Tel: (4) 135 227

Price: Moderate

Wine bar with an extensive selection of unusual and familiar wines

(B6) MURPHY'S

Address: Castle Loop

Tel: (6) 131 352

Price: Moderate

Authentic Irish pub, serving an array of stouts

Cinema listing

Film: Get Johnson
Rating: 18
Time(s): 8 PM, 10 PM everyday
Price: 8 Euros per person

Film: The Hospital
Rating: PG
Time(s): 5 PM, 9 PM
Price: 8 Euros per person

Film: Giraffe
Rating: U
Time(s): 1 PM, 3 PM
Price: 5 Euros per person

Castle

Tours start at 8:30 AM, 10 AM, 1:30 PM, 3 PM.
Tours last 1.5 hours.
Tours cost 10 Euros per person.

Museum

Opens 9 AM – 1 PM, and 2:30 PM – 6 PM.
Entrance costs 10 Euros.
After 4 PM, special reduced admission of 5 Euros.

Post office

Opens 8 AM – Noon and 2:00 – 5:00 PM.

Tower

Tours which go up the whole tower:
Times: 9 AM, Noon, and 3 PM
Price: 15 Euros
Duration: 2.5 hours

Tours of just the lower levels of the tower:
Times: 10 AM, 11 AM, 2 PM, 4 PM
Price: 5 Euros
Duration: 45 minutes

Public transportation: bus

Shown in pink on the map

One bus runs in each direction around the loop shown on the map.

The cost for any single journey on the bus is 1 Euro.

One bus circulates in each direction around the route shown on the map once every 15 minutes.

The first bus starts at 6 AM and the last finishes at 1 AM.

Bus route 1				
Bus stop location	Time (minutes past hour)			
Fountain	--:00	--:15	--:30	--:45
Art Square	--:02	--:17	--:32	--:47
Castle	--:04	--:19	--:34	--:49
Bridge Nine	--:06	--:21	--:36	--:51
Hotel Royal (H3)	--:08	--:23	--:38	--:53
Tourist Info booth	--:10	--:25	--:40	--:55
Shopping area	--:12	--:27	--:42	--:57

Bus route 2				
Bus stop location	Time (minutes past hour)			
Fountain	--:00	--:15	--:30	--:45
Shopping area	--:02	--:17	--:32	--:47
Tourist info booth	--:04	--:19	--:34	--:49
Hotel Royal (H3)	--:06	--:21	--:36	--:51
Bridge Nine	--:08	--:23	--:38	--:53
Castle	--:10	--:25	--:40	--:55
Art Square	--:12	--:27	--:42	--:57

Public transportation: Tram

Shown in yellow on the map

One tram runs between the tower and Arts Square continuously.

The cost for any single journey on the tram is 2 Euros.

One tram completes the whole route, from Arts Square to the tower *and back*, every 20 minutes. The first tram starts 8 AM and the last finishes at 8 PM.

Tram			
Tram stop location	Time		
Art Square	--:00	--:20	--:40
Café Blu (B2)	--:02	--:22	--:42
Red bridge	--:04	--:24	--:44
Main Square	--:06	--:26	--:46
Shopping area	--:08	--:28	--:48
Iron tower	--:10	--:30	--:50
Shopping area	--:12	--:32	--:52
Main Square	--:14	--:34	--:54
Red bridge	--:16	--:36	--:56
Café Blu (B2)	--:18	--:38	--:58
Art Square	--:20	--:40	--:00

INFORMATION FOR DRIVERS

Form W-T

Public parking

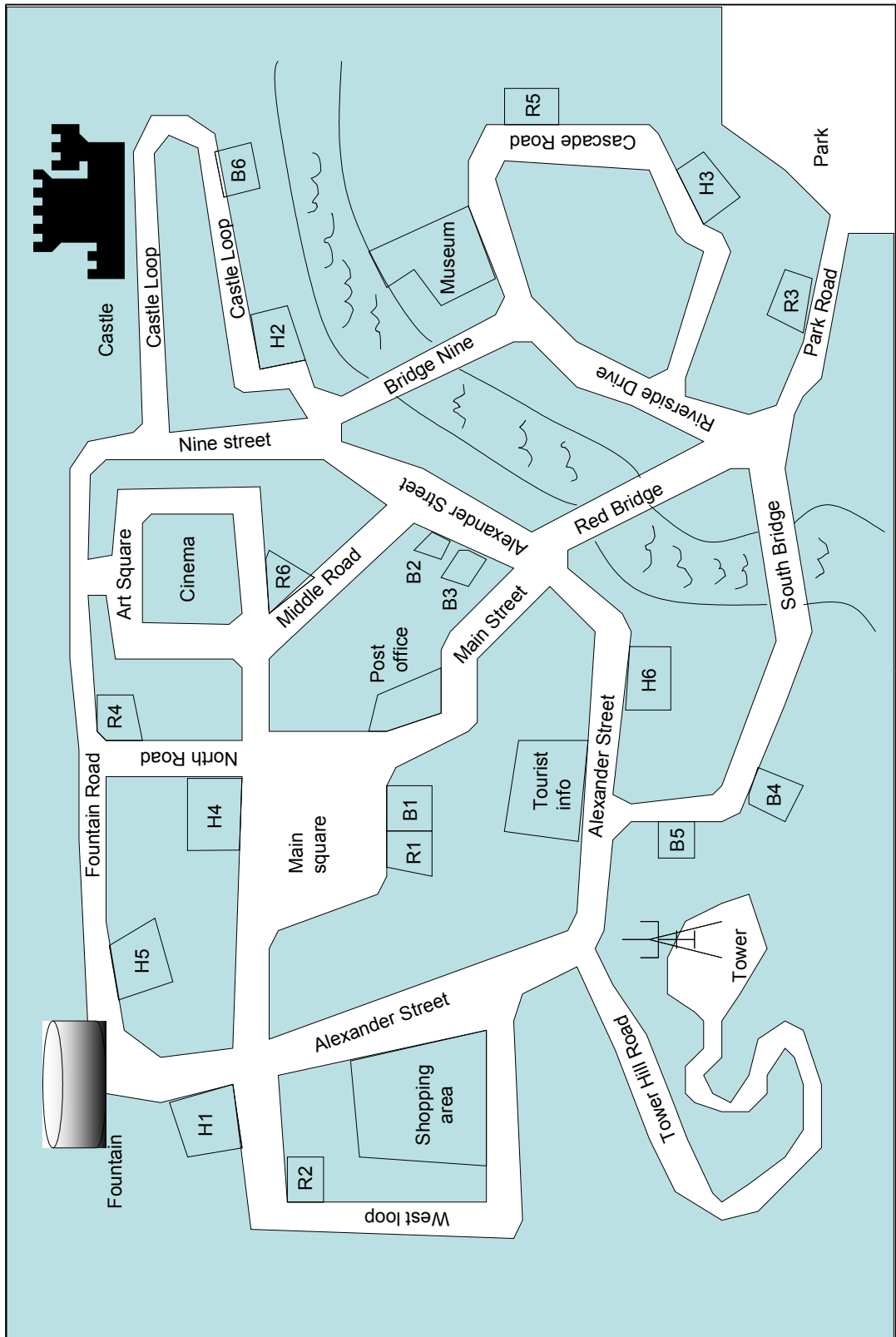
- Underground car parks are below the shopping centre and the museum.
- On the side of the streets in the city centre parking is only allowed for up to 30 minutes.

Traffic restrictions

- One must drive slowly through the park to enter the town centre. This is the only way to get there by car.
- Art Square is closed to car traffic, but it is possible to drive to Main Square from Middle Road.
- Tower Hill Road is closed to car traffic.
- North Road, Main Street to Red Bridge are one way streets. Traffic is southbound only.
- There is often a traffic jam on South Bridge.

Hotels with their own car park are:

- Parking for Hotel Primus and the Art House Hotel is in the garage below the shopping centre.
- The youth hostel has no car park.
- The Royal Hotel, Hotel President, and Alexander Hotel each have their own car parks.



ROLE DESCRIPTION

Form S-R

In this study, you will be asked to complete several tasks, in cooperation with another person, who is also a test subject. You will interact with the other person using a speech recognition simulation. You will hear the other subject when they talk. However, the other subject cannot hear you directly. You will be speaking to a speech recognizer. The speech recognizer will take its best guess at what you've said, and display it on a screen in front of the subject.

When you hear the tic-toc sound, the speech recognizer is working on what you just said. The speech recognizer isn't listening to you when you hear the tic-toc sound. When you hear silence, the speech recognizer is listening (provided that you talk loud enough).

You'll notice that you can "interrupt" the other person. When you talk over them, you'll no longer hear their voice.

In this study, you will be asked to complete a series of tasks – for example, locating a hotel or planning an itinerary.

You will be given the requirements for task, including boxes to fill in and/or maps to annotate. There may be time requirements to satisfy – for example, you may need to complete your errands before a certain time of day, or include an appointment at a specific time.

For each task, you will be given a *basic* street map. The other subject has more information about the town. In order to complete each task, you will need to get the help of the other subject, who has more information about the town, including transportation, locations of shops, and special events.

At the end of each session, you'll be asked for responses to several questions about your experience.

QUESTIONNAIRE

Form S-Q

For each of the following statement, please circle the **one** reaction that best describes the extent to which you agree or disagree with each statement.

1. In this task, I accomplished the goal.

Disagree strongly (1)	Disagree (2)	Disagree somewhat (3)	Neither agree nor disagree (4)	Agree somewhat (5)	Agree (6)	Strongly agree (7)
-----------------------------	-----------------	-----------------------------	--------------------------------------	-----------------------	--------------	-----------------------

2. In this task, I thought the speech recognition was accurate.

Disagree strongly (1)	Disagree (2)	Disagree somewhat (3)	Neither agree nor disagree (4)	Agree somewhat (5)	Agree (6)	Strongly agree (7)
-----------------------------	-----------------	-----------------------------	--------------------------------------	-----------------------	--------------	-----------------------

3. In this task, I found it difficult to communicate because of the speech recognition.

Disagree strongly (1)	Disagree (2)	Disagree somewhat (3)	Neither agree nor disagree (4)	Agree somewhat (5)	Agree (6)	Strongly agree (7)
-----------------------------	-----------------	-----------------------------	--------------------------------------	-----------------------	--------------	-----------------------

4. In this task, I believe the other subject was very helpful.

Disagree strongly (1)	Disagree (2)	Disagree somewhat (3)	Neither agree nor disagree (4)	Agree somewhat (5)	Agree (6)	Strongly agree (7)
-----------------------------	-----------------	-----------------------------	--------------------------------------	-----------------------	--------------	-----------------------

5. In this task, the other subject found using the speech recognition difficult.

Disagree strongly (1)	Disagree (2)	Disagree somewhat (3)	Neither agree nor disagree (4)	Agree somewhat (5)	Agree (6)	Strongly agree (7)
-----------------------------	-----------------	-----------------------------	--------------------------------------	-----------------------	--------------	-----------------------

6. Overall, I was very satisfied with this past task.

Disagree strongly (1)	Disagree (2)	Disagree somewhat (3)	Neither agree nor disagree (4)	Agree somewhat (5)	Agree (6)	Strongly agree (7)
-----------------------------	-----------------	-----------------------------	--------------------------------------	-----------------------	--------------	-----------------------

QUESTIONNAIRE

Form S-Q-a

For each of the following statements please circle the one reaction that best describes the extend to which you agree or disagree with each statement.

1. In this task, I accomplished the goal.

Disagree strongly (1)	Disagree (2)	Disagree somewhat (3)	Neither agree nor disagree (4)	Agree somewhat (5)	Agree (6)	Strongly agree (7)
--------------------------	-----------------	--------------------------	-----------------------------------	-----------------------	--------------	-----------------------

2. In this task, I thought the speech recognition was accurate.

Disagree strongly (1)	Disagree (2)	Disagree somewhat (3)	Neither agree nor disagree (4)	Agree somewhat (5)	Agree (6)	Strongly agree (7)
--------------------------	-----------------	--------------------------	-----------------------------------	-----------------------	--------------	-----------------------

3. In this task, I found it difficult to communicate because of the speech recognition.

Disagree strongly (1)	Disagree (2)	Disagree somewhat (3)	Neither agree nor disagree (4)	Agree somewhat (5)	Agree (6)	Strongly agree (7)
--------------------------	-----------------	--------------------------	-----------------------------------	-----------------------	--------------	-----------------------

4. In this task, I believe the other person was very helpful.

Disagree strongly (1)	Disagree (2)	Disagree somewhat (3)	Neither agree nor disagree (4)	Agree somewhat (5)	Agree (6)	Strongly agree (7)
--------------------------	-----------------	--------------------------	-----------------------------------	-----------------------	--------------	-----------------------

5. In this task, the other person found using the speech recognition difficult.

Disagree strongly (1)	Disagree (2)	Disagree somewhat (3)	Neither agree nor disagree (4)	Agree somewhat (5)	Agree (6)	Strongly agree (7)
--------------------------	-----------------	--------------------------	-----------------------------------	-----------------------	--------------	-----------------------

6. In this task, I found the click-interface very useful.

Disagree strongly (1)	Disagree (2)	Disagree somewhat (3)	Neither agree nor disagree (4)	Agree somewhat (5)	Agree (6)	Strongly agree (7)
--------------------------	-----------------	--------------------------	-----------------------------------	-----------------------	--------------	-----------------------

What did you use it for? _____

7. In this task, I had problems understanding the meaning of the other persons clicking.

Disagree strongly (1)	Disagree (2)	Disagree somewhat (3)	Neither agree nor disagree (4)	Agree somewhat (5)	Agree (6)	Strongly agree (7)
--------------------------	-----------------	--------------------------	-----------------------------------	-----------------------	--------------	-----------------------

8. Overall I was very satisfied with this past task.

Disagree strongly (1)	Disagree (2)	Disagree somewhat (3)	Neither agree nor disagree (4)	Agree somewhat (5)	Agree (6)	Strongly agree (7)
--------------------------	-----------------	--------------------------	-----------------------------------	-----------------------	--------------	-----------------------

ROLE DESCRIPTION

Form W-R

In this study, you will be asked to assist other subjects to complete several tasks. When you speak, the other subject can hear you. However, you can't hear the other subject. When the other subject speaks, they are talking to a speech recognition system. The speech recognizer will take its best guess of what the other subject said, and display it on your screen.

When you don't hear anything, the other subject can hear what you're saying when you talk (as long as you talk loud enough). When you hear the "tic-toc" sound, that means that either the other subject is talking, or the speech recognizer is working on what the other subject just said. The other subject can't hear you when you hear the tic-toc sound.

In this study, the other subject is attempting to plan a series of itineraries.

You will have information the other subject needs to plan their itineraries. Your task is to assist them however you see best.

You will have:

1. A detailed map of the town, including locations of major attractions, shops, and restaurants
2. A host of other information about the town, including a bus timetable, a tram timetable, a listing of restaurants, etc.

The subject also has a map, but their map doesn't have as much detail.

You will have the chance to familiarize yourself with these materials before the start of the sessions.

At the end of each session, you'll be asked for responses to several questions about your experience.

ROLE DESCRIPTION FOR AGENT

Form W-R-a

In this study, you will be asked to assist other participants to complete several tasks. When you speak, the other person can hear you. However, you can't hear the other person. When the other person speaks, they are talking to a speech recognition system. The speech recognizer will take its best guess of what the other person said, and display it on your screen.

When you don't hear anything, the other person can hear what you're saying when you talk (as long as you talk loud enough). When you hear the "tic-toc" sound, that means that either the other person is talking, or the speech recognizer is working on what the other person just said. The other person can't hear you when you hear the tic-toc sound.

In this study, the other person is attempting to plan a series of itineraries.

You will have information the other person needs to plan their itineraries. Your task is to assist them however you see best.

You will have:

- A detailed map of the town, including locations of major attractions, shops, and restaurants
- A host of other information about the town, including a bus timetable, a tram timetable, a listing of restaurants, etc.

The other person also has a map, but their map doesn't have as much detail.

You will have the chance to familiarize yourself with these materials before the start of the sessions.

At the end of each session, you'll be asked for responses to several questions about your experience.

If you find it difficult to understand the other person, because of errors in the speech recognition, you may find it helpful to ask the other participant to use the Point-And-Click interface to assist the communication.

QUESTIONNAIRE

Form W-Q

For each of the following statement, please circle the **one** reaction that best describes the extent to which you agree or disagree with each statement.

1. In this task, I was very helpful to the other subject.

Disagree strongly (1)	Disagree (2)	Disagree somewhat (3)	Neither agree nor disagree (4)	Agree somewhat (5)	Agree (6)	Strongly agree (7)
-----------------------------	-----------------	-----------------------------	--------------------------------------	-----------------------	--------------	-----------------------

2. In this task, I thought the speech recognition was accurate.

Disagree strongly (1)	Disagree (2)	Disagree somewhat (3)	Neither agree nor disagree (4)	Agree somewhat (5)	Agree (6)	Strongly agree (7)
-----------------------------	-----------------	-----------------------------	--------------------------------------	-----------------------	--------------	-----------------------

3. In this task, I found it difficult to communicate because of the speech recognition.

Disagree strongly (1)	Disagree (2)	Disagree somewhat (3)	Neither agree nor disagree (4)	Agree somewhat (5)	Agree (6)	Strongly agree (7)
-----------------------------	-----------------	-----------------------------	--------------------------------------	-----------------------	--------------	-----------------------

4. In this task, the other subject accomplished the goal.

Disagree strongly (1)	Disagree (2)	Disagree somewhat (3)	Neither agree nor disagree (4)	Agree somewhat (5)	Agree (6)	Strongly agree (7)
-----------------------------	-----------------	-----------------------------	--------------------------------------	-----------------------	--------------	-----------------------

5. In this task, the other subject found using the speech recognition difficult.

Disagree strongly (1)	Disagree (2)	Disagree somewhat (3)	Neither agree nor disagree (4)	Agree somewhat (5)	Agree (6)	Strongly agree (7)
-----------------------------	-----------------	-----------------------------	--------------------------------------	-----------------------	--------------	-----------------------

6. Overall, I was very satisfied with this past task.

Disagree strongly (1)	Disagree (2)	Disagree somewhat (3)	Neither agree nor disagree (4)	Agree somewhat (5)	Agree (6)	Strongly agree (7)
-----------------------------	-----------------	-----------------------------	--------------------------------------	-----------------------	--------------	-----------------------

QUESTIONNAIRE

Form W-Q-a

For each of the following statements please circle the one reaction that best describes the extend to which you agree or disagree with each statement.

1. In this task, I accomplished the goal.

Disagree strongly (1)	Disagree (2)	Disagree somewhat (3)	Neither agree nor disagree (4)	Agree somewhat (5)	Agree (6)	Strongly agree (7)
--------------------------	-----------------	--------------------------	-----------------------------------	-----------------------	--------------	-----------------------

2. In this task, I thought the speech recognition was accurate.

Disagree strongly (1)	Disagree (2)	Disagree somewhat (3)	Neither agree nor disagree (4)	Agree somewhat (5)	Agree (6)	Strongly agree (7)
--------------------------	-----------------	--------------------------	-----------------------------------	-----------------------	--------------	-----------------------

3. In this task, I found it difficult to communicate because of the speech recognition.

Disagree strongly (1)	Disagree (2)	Disagree somewhat (3)	Neither agree nor disagree (4)	Agree somewhat (5)	Agree (6)	Strongly agree (7)
--------------------------	-----------------	--------------------------	-----------------------------------	-----------------------	--------------	-----------------------

4. In this task, the other person accomplished the goal.

Disagree strongly (1)	Disagree (2)	Disagree somewhat (3)	Neither agree nor disagree (4)	Agree somewhat (5)	Agree (6)	Strongly agree (7)
--------------------------	-----------------	--------------------------	-----------------------------------	-----------------------	--------------	-----------------------

5. In this task, the other person found using the speech recognition difficult.

Disagree strongly (1)	Disagree (2)	Disagree somewhat (3)	Neither agree nor disagree (4)	Agree somewhat (5)	Agree (6)	Strongly agree (7)
--------------------------	-----------------	--------------------------	-----------------------------------	-----------------------	--------------	-----------------------

6. In this task, I found the click-interface very useful.

Disagree strongly (1)	Disagree (2)	Disagree somewhat (3)	Neither agree nor disagree (4)	Agree somewhat (5)	Agree (6)	Strongly agree (7)
--------------------------	-----------------	--------------------------	-----------------------------------	-----------------------	--------------	-----------------------

What did you use it for? _____

7. In this task, I had problems understanding the meaning of the other persons clicking.

Disagree strongly (1)	Disagree (2)	Disagree somewhat (3)	Neither agree nor disagree (4)	Agree somewhat (5)	Agree (6)	Strongly agree (7)
--------------------------	-----------------	--------------------------	-----------------------------------	-----------------------	--------------	-----------------------

8. In this task, did you ask the user to click on the location, if you had troubles understanding her/him? YES
NO

9. Overall I was very satisfied with this past task.

Disagree strongly (1)	Disagree (2)	Disagree somewhat (3)	Neither agree nor disagree (4)	Agree somewhat (5)	Agree (6)	Strongly agree (7)
--------------------------	-----------------	--------------------------	-----------------------------------	-----------------------	--------------	-----------------------

ROLE DESCRIPTION

Form T-R

In this study, you hear short audio clips of one side of a conversation. The conversation will be happening as you are typing, so speed and accuracy are both important.

In the upper left-hand corner of your screen, you'll see a grey, moving grid. When you see it stop moving, that means the system is waiting for you to enter what was just said.

If you need to hear the sentence again, you can do one of the following things:

1. If you haven't typed anything yet, you can just type a space (i.e., press the spacebar) and press enter.
2. Or, you can click on the button that says "Replay last utterance." After you do that, you'll need to click on the AIM window again to start typing.

When you've typed what you hear, press enter.

If you type something and press enter *before* you see the red line, it will store what you typed until the red line appears. You shouldn't type it a second time.

If the recording contains only silence (i.e., no speech), just type a full stop (.) and press enter.

If the recording contains any words that you don't know how to spell, or can't understand, take your best guess. If you can't understand what the person is saying, it's not important if you're sure you've got it right. There is no way to enter "I don't know"!

It's not important whether you use all lower case or capital letters – do whatever is easiest for you. It's also not important whether you use periods, commas, semi-colons, etc. – again, do whatever is easiest. However, you *should* use apostrophes, like in *can't* or *shouldn't*.

Sometimes people say things that aren't really words. When you hear these, try to use the following "talking words":

- uh
- ah
- er
- um
- oh

You'll be given time to familiarize yourself with the system before a "real" conversation starts.

Appendix B: Wizard transcription guidelines

In general, the goal of transcription is to capture exactly what is said -- *not* to make a speaker's meaning clear. Thus, transcribers should transcribe *exactly what they hear*. For example, do not correct grammatical mistakes, do not remove hesitation words, and do not modify what is said to make its meaning clearer. Attention to detail is absolutely crucial!

Use all lower-case letters, even for proper nouns, for individual letters, and for "i".

Each convention below has been copied directly from the LDC RT-03 manual, including LDC section numbers.

Use British English spellings. Check spellings at www.dictionary.com.

- [3.2.1.2] **Spelling:** Transcribers use standard orthography, word segmentation and word spelling. All files must be spell-checked after transcription is complete. When in doubt about the spelling of a word or name, annotators consult a standard reference, like an online or paper dictionary, world atlas or news website. *Also, please consult the provided materials, including the map and "Town information" for spellings.*
- [3.2.1.3] **Contractions:** Annotators limit their use of contractions to those that exist in standard written English, and of course only when a contraction is actually produced by the speaker. Annotators must take care to transcribe exactly what the speaker says. The table below, while not comprehensive, shows some examples of how to transcribe common contractions.

Complete Form	Spoken As	Transcribed As	Incorrect
I have	<i>I've</i>	I've	
cannot	<i>can't</i>	can't	
will not	<i>won't</i>	won't	
you have	<i>you've</i>	you've	
could not	<i>couldn't</i>	couldn't	
should have	<i>should've</i>	should've	should of, shoulda
would have	<i>would've</i>	would've	would of, woulda
it is	<i>it's</i>	it's	its
its (possessive)	<i>its</i>	its	it's
Marvin (possessive)	<i>Marvin's</i>	Marvin's	
Marvin is	<i>Marvin's</i>	Marvin's	
Marvin has	<i>Marvin's</i>	Marvin's	
going to	<i>gonna</i>	going to	gonna
want to	<i>wanna</i>	want to	wanna
got to	<i>gotta</i>	got to	gotta

Note: Annotators should take care to avoid the common mistakes of transposing possessive its for contraction it's (it is), possessive your for the contraction you're (you are), and their (possessive), they're (they are) and there.

Annotators should transcribe exactly what they hear using standard orthography. If a speaker uses a contraction, the word is transcribed as contracted: they're, won't, isn't, don't and so on. If the speaker uses a complete form, the annotator should transcribe what is heard: they are, is not and so on.

For non-standard contractions like "gonna" and "wanna" annotators should spell out the entire word: going to, want to.

- [3.2.1.4] **Numbers:** All numerals are written out as complete words. Hyphenation is used for numbers between twenty-one and ninety-nine only.
 - ◆ twenty-two
 - ◆ nineteen ninety-five

- ◆ seven thousand two hundred seventy-five
- ◆ nineteen oh nine
- [3.2.1.5] **Words and compounds:** In general, annotators should be conservative about use of hyphens. For instance:
 - ◆ an overly complicated analysis
 - ◆ *not* an overly-complicated analysis

However, in some cases, a hyphen is required:

- ◆ anti-nuclear protests
- ◆ *not* anti nuclear protests

Compounds can be tricky. When in doubt, annotators should consult a dictionary and talk to their language team leader.

- [3.2.2.2] **Filled pauses and hesitation sounds:** Filled pauses are non-lexemes (non-words) that speakers employ to indicate hesitation or to maintain control of a conversation while thinking of what to say next. Each language has a limited set of filled pauses that speakers can employ. Annotators use the standardized spellings shown in the table below for filled pauses. The spelling of filled pauses is not altered to reflect how the speaker pronounces the word (e.g., typing AH for a loud "ah" or ummmm for a long "um".) For English, this set includes ah, eh, er, uh, um.

All filled pauses are indicated with a % sign preceding the word.

English Filled Pauses	Arabic Filled Pauses	Chinese Filled Pauses
%ah	%ah	%阿
%eh	%E	%呃
%er	%M	%唔
%uh	%uh	
%um	%hm	
	%hum	

- [3.2.2.3] **Partial words:** When a speaker breaks off in the middle of the word, annotators transcribe as much of the word as can be made out. A single dash - is used to indicate point at which word was broken off.
 - ◆ yes, absolu- absolutely.
- [3.2.2.4] **Restarts:** Speaker restarts are indicated with double dash --. Annotators use this convention for cases where a speaker stops short, cutting him/herself off before continuing with the utterance.
 - ◆ i thought he -- i thought he was there
 - ◆ the thi- -- the thing we're worried about is
- [3.2.4.1] **Hard-to-understand sections:** Sometimes an audio file will contain a section of speech that is difficult or impossible to understand. In these cases, annotators use double parentheses (()) to mark the region of difficulty.

Sometimes it is possible to take a guess about the speaker's words. In these cases,

annotators transcribe what they think they hear and surround the stretch of uncertain transcription with double parentheses:

- ◆ and she told me that ((i should just leave))

If an annotator is truly mystified and can't at all make out what the speaker is saying, s/he uses empty double parentheses to surround the untranscribed region. Where possible, this untranscribed region gets its own timestamp, e.g.:

- ◆ (())

- [3.2.4.2] **Idiosyncratic words:** Occasionally a speaker will make up a new word on the spot. These are not the same as slang words; they're words that are unique to the speaker in that conversation. If annotators encounter an idiosyncratic word, they should transcribe it to the best of their ability and mark it with an asterisk *. For instance:

- ◆ do you dress like a *schlump yet
- ◆ why she said *drr i don't know

- [3.2.4.5] **Interjections:** The following standardized spellings are used to transcribe interjections. Interjections do not require any special symbol.

English Interjections

ach	huh-uh	oh	whew
duh	hm	okay	whoops
eee	jeepers	oof	woo-hoo
ew	jeez	ooh	yay
ha	mm	uh-huh	yeah
hee	mhm	uh-oh	yep
huh	nah	whoa	yup

The following conventions have been added to the LDC conventions. They are intended to extend and be compatible with the LDC conventions.

- **Self-speech:** When the speaker is clearly talking to him or her-self, and the speech is not directly intended for the listener, transcribe the "self-speech" in the tags [self-speech] ... [/self-speech].

- ◆ [self-speech] the castle the castle [/self-speech]
right so the castle is on castle loop

When it's not clear whether speech is directed at the listener or not, err on the side of not using the [self-speech] tags -- i.e., "When in doubt, leave them out." In general, hesitation words on their own should not be included in [self-speech] tags.³

- **End-pointer errors:** It is possible that the beginning or end of the utterance has speech in which the speaker has been "interrupted" - i.e., the recording started a little after the speaker started, or just before the speaker ended. These interruptions are called "End-pointer errors" and should be noted in two ways.

First, in addition to the transcription column, there are two other columns, "ep-error-start" and "ep-error-end." These columns are normally both blank. If there is an interruption caused by the end-pointer at the beginning or end of the utterance, note that by entering "true" in the appropriate column.

Second, the word fragment you hear should be included in the transcription. For words cut off at the beginning of the utterance, put the dash before the word; for word fragments cut off at the end of the utterance, put the dash at the end of the word.

³ The LDC convention uses <side-speech> ... </side-speech> - however, this causes problems with ANVIL, so it has been changed here.

Appendix C: Converting from HTK WAV to RIFF WAV

HTK audio is stored as WAV (PCM) data. Because it includes HTK-specific header information, it is not immediately playable by most audio tools, which expect RIFF WAV data.

HTK audio is stored as WAV data with the following characteristics:

- 16 kHz
- Mono
- 16-bit
- Big-endian

You can use command line-tools to convert HTK-style audio to RIFF. One example is sox, which can be obtained from <http://sox.sourceforge.net/>

```
sox -t .raw -r 16000 -c 1 -w -s -x in.wav out.wav
```

Note that this treats the header as audio, so there is a short "pop" at the beginning of the sound file. A second option (which removes the header) is to use the HTK tool HCopy:

```
HCOPY -C config.code inHTK.wav outRAW.wav
```

with config.code file being

```
SOURCEFORMAT = HTK  
TARGETFORMAT = NOHEAD  
NATURALBYTEORDER = TRUE
```

Another option is to open HTK audio directly in a hi-quality sound editor like Adobe Audition (formerly called CoolEdit). In Adobe Audition, select the following options when opening:

- 16 kHz
- Mono
- 16-bit
- 16-bit Motorola PCM (MSB,LSB)
- 0 bit offset to input data

Note the header will again be treated as audio & there will be a short initial pop.