

The *SACTI-1* Corpus:
Guide for Research Users

Jason D. Williams, Steve Young
{jdw30, sjy}@eng.cam.ac.uk

22 February 2005 (Version 1.0)

University of Cambridge
Department of Engineering
Technical Report: CUED/F-INFENG/TR482

Table of contents

INTRODUCTION	4
AUDIENCE AND SCOPE	4
CORPUS OVERVIEW	4
ACKNOWLEDGEMENTS	4
CORPUS LICENSE	4
DATA COLLECTION ENVIRONMENT	5
OVERVIEW OF THE SIMULATED ASR-CHANNEL	5
TURN-TAKING MODEL IN THE SIMULATED ASR CHANNEL	6
FULL-AUDIO DIALOGS	6
DATA COLLECTION METHOD	7
CORPUS CONTENTS	8
SUMMARY	8
FILE NAMING AND ORGANIZATION	9
TRANSCRIPTION	9
TRANSCRIPTION OF THE SIMULATED ASR CHANNEL DIALOGUES	9
<i>User utterance transcription in the Simulated ASR Channel</i>	9
<i>Wizard utterance transcription in the Simulated ASR Channel</i>	9
TRANSCRIPTION IN THE FULL-AUDIO DIALOGUES	10
AUTOMATIC TURN ANNOTATION	10
AUTOMATIC TURN ANNOTATION IN THE SIMULATED ASR DIALOGUES	10
AUTOMATIC TURN ANNOTATION IN THE FULL-AUDIO DIALOGUES	10
WIZARD UNDERSTANDING STATUS ANNOTATION	11
GROUNDING ACT ANNOTATION	11
TASK-COMPLETION ANNOTATION	16
ANVIL SCHEMA DETAILS	16
SIMULATED ASR DIALOGUES FILE DETAILS	17
<i>“states” track</i>	17
<i>“turns” track</i>	18
FULL-AUDIO DIALOGUES FILE DETAILS	18
<i>“userUtts” track and “wizardUtts” track</i>	18
<i>“turnsFullAudio” track</i>	19
DIALOGUE PROPERTIES SET	19
EXAMPLE SIMULATED ASR CHANNEL ANVIL FILE	20
REFERENCES	22
APPENDIX A: EXPERIMENTATION DOCUMENTS	24
APPENDIX B: WIZARD TRANSCRIPTION GUIDELINES	66
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-1* dialogue corpus. *SACTI* stands for *Simulated ASR Channel, Tourist Information*.

Audience and scope

This document is intended as a stand-alone “manual” for researchers interested in using the SACTI-1 corpus for research endeavours.

Work published in conference proceedings has discussed the “Simulated ASR Channel” collection framework [1], and analyzed the corpus for various conversational phenomena [2]. Further work is on-going. This document is intended as a stand-alone document; as such, it contains some overlap with these two works.

Note that this document is not intended to motivate the collection framework, nor explain or justify the ASR confusion simulation; for details on this topic, see [1]. Further, this document is also not intended to analyze or draw conclusions from the contents of the SACTI-1 corpus; for details on this topic, see [2].

Corpus overview

The SACTI-1 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. The first (and larger) part of the corpus contains human-human dialogues in a “simulated automated speech recognition (ASR) channel.” This portion of the data is intended to be useful for training statistical components. The second (smaller) part contains human-human dialogues in a full-audio channel. This portion of the data is intended to be useful by providing a baseline – i.e., allowing direct comparisons of “natural” human-human conversation vis-à-vis the simulated ASR channel.

Acknowledgements

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

- Michael Kipp, for making his ANVIL tool available.
- Carol Lions at Human Resources, Cambridge, for staffing wizards.
- Karl Weilhammer, for helpful discussions regarding transcription conventions and tools.
- Matt Stuttle and Karl Weilhammer for assistance with conducting the tests

Portions of this work were supported by the European Union Framework 6 TALK Project (507802).

Corpus license

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

Data collection environment

This section first describes the data collected for the simulated ASR channel dialogs, then the full-audio dialogs.

Overview of the simulated ASR-channel¹

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

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. 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.

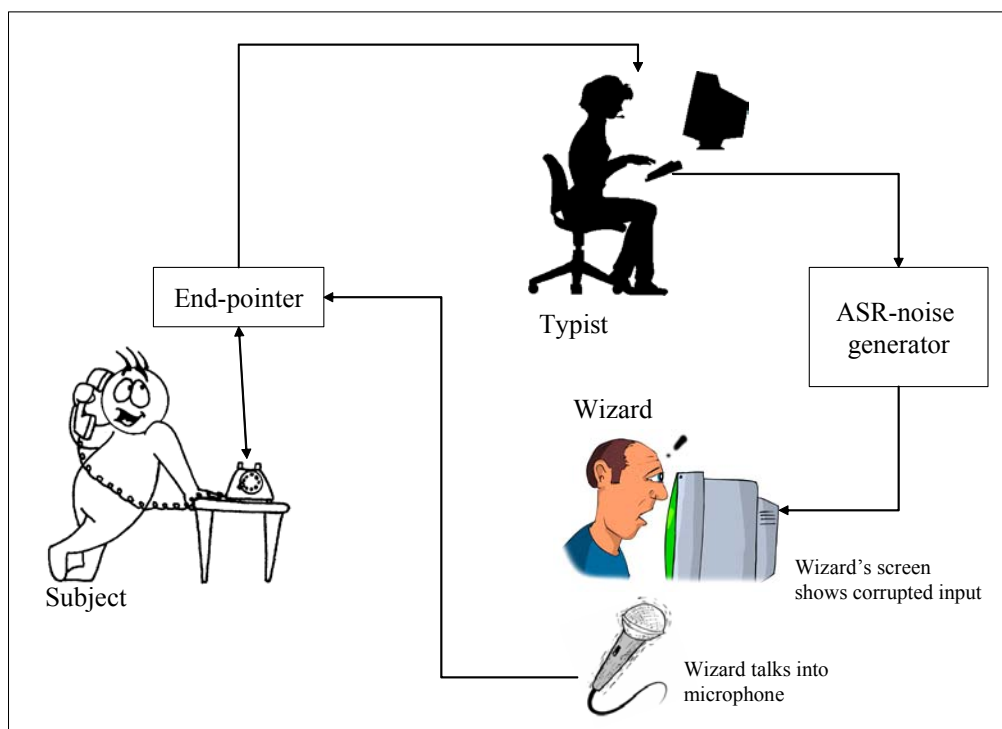


Figure 1: High-level view of collection framework

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.

When the system is busy and not listening to a participant, they hear a “tick-tock” sound. A turn-taking model patterned after typical HC 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.

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

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).

Turn-taking model in the simulated ASR channel

Internally there are 5 system states:

- **SILENCE:** Either participant can begin speaking; both hear silence.
- **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.
- **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.
- **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.
- **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.

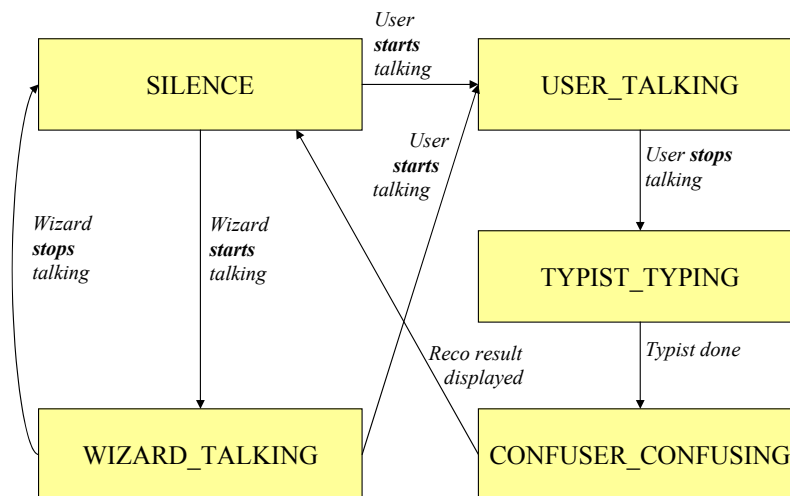


Figure 2: Transition diagram for state machine (turn-taking model)

Full-audio dialogs

Unlike the simulated ASR channel dialogues, the full-audio dialogues use a full-duplex audio interface – effectively a high-fidelity telephone – and the participants can hear each other directly

and instantly. No computer display is used. To mimic a telephone environment, the participants can hear their own voice in addition to the voice of the other participant.

Since there is no enforced turn-taking, full audio recordings of both sides of the conversations are made.

In all other respects the full-audio dialogs were conducted using the same methodology as the simulated ASR channel dialogs.

Data collection method

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

Tests were structured in three-hours blocks, during which one wizard interacted with three users. Each user was given 4 different tasks; thus each wizard engaged in 12 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 24 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 same typist.

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, 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 said “I’m looking for your honest reactions to these questions. Don’t feel you’re grading the system or me – your honest feedback is most important.”

Corpus contents

Summary

The corpus includes the following dialogues:

Channel type	ASR target	Wizard ID	User ID	Number of Dialogues	Task IDs	Collection Date
Sim-ASR	Med	001	001	4	1-1, 1-2, 1-3, 1-4	6-Feb-04
Sim-ASR	Med	001	002	4	2-1, 2-2, 2-3, 2-4	6-Feb-04
Sim-ASR	Med	001	003	4	3-1, 3-2, 3-3, 3-4	6-Feb-04
Sim-ASR	Med	002	004	4	4-1, 4-2, 4-3, 4-4	6-Feb-04
Sim-ASR	Med	002	005	4	5-1, 5-2, 5-3, 5-4	6-Feb-04
Sim-ASR	Med	002	006	4	6-1, 6-2, 6-3, 6-4	6-Feb-04
Sim-ASR	Low	003	007	4	1-1, 1-2, 1-3, 1-4	10-Feb-04
Sim-ASR	Low	003	008	4	2-1, 2-2, 2-3, 2-4	10-Feb-04
Sim-ASR	Low	003	009	4	3-1, 3-2, 3-3, 3-4	10-Feb-04
Sim-ASR	Low	004	010	4	4-1, 4-2, 4-3, 4-4	10-Feb-04
Sim-ASR	Low	004	011	4	5-1, 5-2, 5-3, 5-4	10-Feb-04
Sim-ASR	Low	004	012	4	6-1, 6-2, 6-3, 6-4	10-Feb-04
Sim-ASR	Hi	005	013	4	1-1, 1-2, 1-3, 1-4	12-Feb-04
Sim-ASR	Hi	005	014	4	2-1, 2-2, 2-3, 2-4	12-Feb-04
Sim-ASR	Hi	005	015	4	3-1, 3-2, 3-3, 3-4	12-Feb-04
Sim-ASR	Hi	006	016	4	4-1, 4-2, 4-3, 4-4	12-Feb-04
Sim-ASR	Hi	006	017	4	5-1, 5-2, 5-3, 5-4	12-Feb-04
Sim-ASR	Hi	006	018	4	6-1, 6-2, 6-3, 6-4	12-Feb-04
Sim-ASR	Med	007	019	4	1-1, 1-2, 1-3, 1-4	17-Feb-04
Sim-ASR	Med	007	020	4	2-1, 2-2, 2-3, 2-4	17-Feb-04
Sim-ASR	Med	007	021	4	3-1, 3-2, 3-3, 3-4	17-Feb-04
Sim-ASR	Med	008	022	4	4-1, 4-2, 4-3, 4-4	17-Feb-04
Sim-ASR	Med	008	023	4	5-1, 5-2, 5-3, 5-4	17-Feb-04
Sim-ASR	Med	008	024	4	6-1, 6-2, 6-3, 6-4	17-Feb-04
Sim-ASR	Low	009	025	4	1-1, 1-2, 1-3, 1-4	18-Feb-04
Sim-ASR	Low	009	026	4	2-1, 2-2, 2-3, 2-4	18-Feb-04
Sim-ASR	Low	009	027	4	3-1, 3-2, 3-3, 3-4	18-Feb-04
Sim-ASR	Low	010	028	4	4-1, 4-2, 4-3, 4-4	18-Feb-04
Sim-ASR	Low	010	029	4	5-1, 5-2, 5-3, 5-4	18-Feb-04
Sim-ASR	Low	010	030	4	6-1, 6-2, 6-3, 6-4	18-Feb-04
Sim-ASR	None	011	031	4	1-1, 1-2, 1-3, 1-4	19-Feb-04
Sim-ASR	None	011	032	4	2-1, 2-2, 2-3, 2-4	19-Feb-04
Sim-ASR	None	011	033	4	3-1, 3-2, 3-3, 3-4	19-Feb-04
Sim-ASR	None	012	034	4	4-1, 4-2, 4-3, 4-4	19-Feb-04
Sim-ASR	None	012	035	4	5-1, 5-2, 5-3, 5-4	19-Feb-04
Sim-ASR	None	012	036	4	6-1, 6-2, 6-3, 6-4	19-Feb-04
Full-audio	FullAudio	013	037	4	1-1, 1-2, 1-3, 1-4	2-Dec-03
Full-audio	FullAudio	013	038	4	2-1, 2-2, 2-3, 2-4	2-Dec-03
Full-audio	FullAudio	013	039	4	3-1, 3-2, 3-3, 3-4	2-Dec-03
Full-audio	FullAudio	014	040	4	4-1, 4-2, 4-3, 4-4	2-Dec-03
Full-audio	FullAudio	014	041	4	5-1, 5-2, 5-3, 5-4	2-Dec-03
Full-audio	FullAudio	014	042	4	6-1, 6-2, 6-3, 6-4	2-Dec-03

Table 1: Summary of corpus contents

File naming and organization

Dialogue files are represented in the format WWW-UUU-D (for example 002-006-2), where:

- WWW = the ID of the wizard, ranging from:
 - 001 to 012 for simulated ASR channel dialogues, and
 - 013 and 014 for full-audio dialogues
- UUU = the ID for the user, ranging from
 - 001 to 036 for simulated ASR channel dialogues, and
 - 037 to 042 for full-audio dialogues
- D = the ID of the dialog between that user and that wizard, ranging from
 - 1 to 4 for all dialogues.

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

Each dialog includes the following files:

- WWW-UUU-D.anvil: The ANVIL XML file, including transcriptions (described below).
- WWW-UUU-D.wav: A 2-channel (stereo) RIFF WAV audio file with both sides of the conversation
- WWW-UUU-D.mov: A 1-channel (mono) QuickTime movie with both sides of the conversation²

The WAV to QuickTime conversation was performed with [11].

For information on converting HTK WAV to RIFF WAV, see Appendix C.

Transcription

Transcription of the Simulated ASR Channel Dialogues

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 dialogs. Thus the wizard utterances could be transcribed for a richer set of phenomena.

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]

We are concerned here with the orthographic representation of the wizard’s speech, corresponding 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.

Transcription in the Full-Audio dialogues

The full-audio dialogues (both the wizard and user sides) were transcribed using PRAAT off-line, after the conclusion of the dialogues. For the full-audio dialogues, no notion of “utterance” was imposed by the interface, so the transcriptionist first segmented the wizard and user tracks into utterances – i.e., segments of continuous speech. The transcription conventions for the wizard side of the simulated ASR dialogues were used for both sides of the full-audio dialogues.

Automatic turn annotation

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.

Automatic turn annotation in the Simulated ASR Dialogues

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.

Automatic turn annotation in the Full-Audio Dialogues

In the full-audio dialogues, the participants can speak at the same time, so we must use a slightly more complex definition of “turn”. In the full-audio dialogues we define speaker A’s turn as beginning when speaker A is speaking when the other is silent, and as ending when that speaker A stops speaking, and isn’t the next to speak. Note that speaker A’s turn may cover multiple utterances from speaker A, provided that speaker B does not speak during the gaps between speaker A’s utterances.

This definition allows a speaker to use back-channel without taking a turn.

We believe this definition is comparable to that used in the simulated ASR dialogues and thus allows a reasonable quantitative comparison of turn-taking rates.

Wizard understanding status annotation

At present, the “Wizard understanding status annotation” has been done only for the Simulated ASR Channel, not the Full-Audio channel.

Each wizard turn was annotated to indicate whether the intention expressed in the *previous* user turn was understood. Both the recognition result and the wizard's actions are taken into account when assessing understanding status. If the meaning of the utterance was clearly obscured by the end-pointer, this should be considered in classification of this utterance.

The following categories are used:

Category	Description
<i>Full</i>	The full intention of the user was understood or inferred. For example, if the wizard seeks to confirm an intention, and that intention is correct, then it should be labelled <i>Full</i> . The wizard’s response may include a little extra information than what the user requested.
<i>Partial</i>	A portion of the user's intention was understood or inferred, and nothing was misunderstood. Also, if the wizard provided a “shotgun response” – i.e., lots of information which happens to include the requested information – this should be annotated as <i>Partial</i> .
<i>Non</i>	No part of the user's intention was understood. (Individual words may have been communicated by the recognizer, but the wizard could not form a plausible hypothesis based on them.)
<i>Mis</i>	At least one aspect of the user's intention was understood incorrectly. For example, if the wizard seeks to confirm an intention, and that intention isn't correct, then it should be labelled <i>Mis</i> . If there were no ASR errors yet the meaning was still misunderstood, that misunderstanding should still be annotated. Finally, if the wizard's interpretation or reaction shows a misunderstanding, but then goes on to give an answer to the user's question, that should still be labelled as <i>Mis</i> . For example, “What are the films on at the cinema?” / “No, there is only one cinema, and it's playing X, Y, and Z” should be labelled <i>Mis</i> .
<i>NA-Wizard</i>	Used when the wizard speaks first in a dialog, or when it cannot be inferred from the data whether the wizard understood the user’s intention.

Table 2: Listing of Wizard Understanding Status categories.
Each wizard turn is classified into one of these categories.

Grounding act annotation

At time of printing, the “Grounding act annotation” has been done only for the Simulated ASR Channel, not the Full-Audio channel.

We were interested to study the grounding process at a relatively low level. We sought to annotate each wizard and user turn to indicate what grounding behaviors were expressed.

There are a cornucopia of schemes for annotating “Dialog Acts” in spoken conversation. We examined: schemes used in similar experiments by Skantze [13]; Augmented C-STAR [14]; DAMSL [15], HCRC Map Task Moves [16], Verbmobil [17], ATR [18], and Traum [19].

Traum’s Conversation Acts were closest to our needs, and broadly inspired our set. As we were interested in examining actions on a more granular level than Conversations Acts show, we created the following set of “Grounding Tags.”

Tag	Description	Example(s)	Notes
<i>Inform</i>	Statement or repetition of fact or desire.	U: "I'd like to go to the tower." W: "There are three hotels that are near the main square." W: "I don't have any other restaurants."	<ul style="list-style-type: none"> "No" answers to an offer are an <i>Inform</i>. E.g., "Do you want to hear about the other restaurants?" / "No. Can you tell me about cheap hotels?" is <i>Inform + Request</i>.
<i>Request</i>	Question or request (possibly repeated) which invites the other party to make a contribution.	W: "What do you want to do?" U: "Does bus one go there?" U: "How much does that cost?" W: "Can I help with anything else?"	<ul style="list-style-type: none"> Asking for elaboration on a query, and providing an elaborated query, are both <i>Requests</i>. I.e., suppose agent-1 asks a question, and agent-2 understand the question but needs agent-1 to be more specific so asks for more information, then agent-1 provides a more elaborated query. In this example, all turns are annotated as <i>Request</i>. However, if agent-1's response does not provide an elaborated query but rather provides the answer to Agent-2's question, then agent-1's response is annotated as <i>Inform</i>.
<i>GreetingFarewell</i>	Conversation opening, closing, and small talk. The wizard's opening may invite an initial response from the user.	W: "Oh, hi how can I help?" U: "Oh, hi there." U: "That's it. End Task."	<ul style="list-style-type: none"> "Hi" at the beginning of an utterance is annotated as <i>GreetingFarewell</i> (The first) "How can I help?" without a "hi" or "hello" is annotated as <i>GreetingFarewell</i>. Closing utterances like "That's great, thank you very much" and "you're welcome - would you like anything else?" and "No thanks" and "Do you have any other questions?" are all annotated as <i>GreetingFarewell</i>.

Tag	Description	Example(s)	Notes
<i>ExplAck</i>	Explicit show of (perceived!) understanding—i.e., “I understand you.”	W: “Right.” U: “Ok.”	<ul style="list-style-type: none"> • “Yes” which answers a yes/no question is not annotated as <i>ExplAck</i> but rather as <i>Inform</i>. • “Yes” at the beginning of an utterance which is not responding to a yes/no question is annotated as <i>ExplAck</i>. • “No” which is not responding to a yes/no question but which means “I can’t do that” is not annotated as <i>ExplAck</i>. For example, “Can you tell me about the bar?” / “Yes, it’s upmarket and rather expensive” - the second turn is annotated as <i>ExplAck</i> + <i>Inform</i>; but “Can you tell me where I can get a pizza in town?” / “No, there isn’t a pizza restaurant in town” - the second turn is annotated as <i>Inform</i>. • “And” at the beginning of an utterance is not annotated as <i>ExplAck</i> • “Thank you” and “you’re welcome” are annotated as <i>ExplAck</i>.
<i>ReqAck</i>	Explicit request for communication of the other party’s understanding.	W: “Does that make sense?” W: “Does that answer your question?” U: “Did that come through ok?”	<ul style="list-style-type: none"> • <i>ReqAck</i> is used to annotate “are you there” queries—e.g., in the middle of a dialog, “Hello?” • “Would you like me to repeat that?” is annotated as <i>ReqAck</i>.
<i>UnsolicitedAffirm</i>	An unsolicited affirmation that the other party has understood the speaker’s past utterance correctly (i.e., not preceded by <i>ReqAck</i>).	U: “Yes, that’s right.” U: “Yes, I want to go to the castle.”	--
<i>RespondAffirm</i>	An explicit positive response to <i>ReqAck</i> .	U: “Yes, that’s right.” U: “Yes, I want to go to the castle.”	--
<i>RespondNegate</i>	An explicit negative response to <i>ReqAck</i> .	U: “No, that’s not right.” U: “No.”	<ul style="list-style-type: none"> • The sequence “Would you like me to repeat that” / “yes” is annotated as <i>ReqAck</i> / <i>RespondNegate</i>.

Tag	Description	Example(s)	Notes
<i>RejectOther</i>	Rejection of the other party's interpretation of the speaker's intention without a preceding <i>ReqAck</i> —i.e., an uninitiated show that the other party hasn't understood the speaker.	U: "No, that's not right." U: "You've got it wrong."	--
<i>DisAck</i>	Explicit display of non-understanding—i.e., show of absence of any contribution from other agent's last turn.	W: "Uhh... I don't understand." W: "Nope, didn't catch that."	<ul style="list-style-type: none"> • "I didn't understand what you said" and "I don't follow what you mean" are both annotated as <i>DisAck</i>. • "What do you mean?" is annotated as <i>DisAck</i> (and not as <i>ReqRepeat</i>). • When "sorry" implies the intention is "I don't know what you mean/said", it is annotated as <i>DisAck</i>. For example, "Sorry, can you repeat that?" is annotated as <i>DisAck</i> + <i>ReqRepeat</i>.
<i>StateInterp</i>	A statement of belief or hypothesis in the desire or intention of the other agent.	W: "So, you want prices of hotels near the main square." W: "... prices of hotels near the main square." W: "I guess you want prices of hotels near the main square."	<ul style="list-style-type: none"> • The <i>StateInterp</i> tag may refer to any earlier turn. • The <i>StateInterp</i> tag may refer to one piece of information – it need not refer to the other party's complete desire/intention. • <i>StateInterp</i> is used when the statement of the other agent's desires/intentions is offset from the rest of the speaker's turn. For example, "Ok, near the main square... there is one hotel near the main square" is annotated as <i>ExplAck</i> + <i>StateInterp</i> + <i>Inform</i>, whereas "Ok, there is one hotel near the main square" is annotated as <i>ExplAck</i> + <i>Inform</i>.
<i>ReqRepeat</i>	An invitation for the other agent to state their desire / intention again.	W: "Can you say that again please." U: "Can you repeat please?"	<ul style="list-style-type: none"> • Re-asking the same question is not annotated as <i>ReqRepeat</i>. <i>ReqRepeat</i> is used when one speaker invites the other to say something "again," "a second time", "another time", etc. • Requests to "rephrase" are annotated as <i>ReqRepeat</i>. For example, "Could you rephrase that?" • <i>ReqRepeat</i> refers only to recent items. Requests to repeat more distant information from earlier in the dialog are annotated as <i>Request</i>.

Tag	Description	Example(s)	Notes
<i>HoldFloor</i>	A sound or request intended to keep the floor while not making a contribution.	W: “Uhh, so... just a sec.” U: “Oh... wait...”	<ul style="list-style-type: none"> • <i>HoldFloor</i> is used when there is a clear request to wait like “I’m looking” or “just a minute” etc. • Turns which begin with dysfluencies are not annotated with <i>HoldFloor</i>. • <i>HoldFloor</i> can be used in a sequence of grounding tags as well as used to annotate a whole turn • Singing while looking for a piece of information is annotated as <i>HoldFloor</i>.
<i>IncompleteUnknown</i>	A complete turn which has no interpretation within the above tags. For example, this tag can be used for extreme end-pointing errors.	W: “Uh, so...” [end-pointer then interrupts wizard] U: “[noise]” [end-pointer error]	<ul style="list-style-type: none"> • Turns which include just dysfluency are annotated as <i>IncompleteUnknown</i>

Table 3: “Grounding Tags” used in this corpus.
Each wizard & each user turn was annotated with an ordered list of one or more of these tags.

Each turn is annotated as a sequence of one or more of the tags above. Where possible, the surface order of the intentions is preserved. For example:

Example Transcript	Grounding tags annotation
“So, you want to go to the castle?”	StateInterp + ReqAck
“So, you want to go from the castle to the tower, let me see... you should take bus number one.”	StateInterp + Inform
“No, I don't understand - can you say again?”	DisAck + ReqRepeat
“It says 'pizza street.' I don't think that's right.”	StateInterp + DisAck

Table 4: Example grounding tag annotations

Finally, self-corrections are not annotated – i.e., only the *corrected* portion of a self-correction are annotated.

Task-completion annotation

Included with each dialog 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)

ANVIL schema details

Each dialogue is stored as an ANVIL-format XML file.

The simulated ASR dialogues are defined in the XML file “woz-anvil-spec.xml”. Each file (dialogue) contains 2 ANVIL “tracks” – one for “states” one for “turns” – and 1 ANVIL “set”.

The full-audio dialogues are defined in the XML file “fullaudio-anvil-spec.xml”. Each file (dialogue) contains 3 ANVIL “tracks” – one track of “utterances” for each participant and one showing “turns” – and 1 ANVIL “set”.

Simulated ASR dialogues file details

“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. A second track called “turns” is used to show the progression of user and wizard turns.

The ANVIL track “states” contains the raw output of the state machine. 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	No handler in transition hash for state=USER_TALKING, id=tm.typist, cmd=silence, msg=silence

	traffic. This does not indicate that the participants experienced a problem with the simulation.	
next	Indicates the <i>type</i> of the next state.	Same as <i>type</i> , above.

Table 5: 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
prevUserTurnUnderstanding	Wizard understanding status annotation	<i>For wizard turns:</i> Full, Partial, Non, Mis, NA-Wizard <i>For all user turns:</i> NA-User
groundingAct01 ... groundingAct09	Ordered sequence of “Grounding Tags”	Inform, Request, GreetingFarewell, ExplAck, ReqAck, UnsolicitedAffirm, RespondAffirm, RespondNegate, RejectOther, DisAck, StateInterp, ReqRepeat, HoldFloor, IncompleteUnknown

Table 6: Properties of the *turns* ANVIL track in the Simulated-ASR dialogues

Full-audio dialogues file details

“userUtts” track and “wizardUtts” track

In the full-audio dialogues, the Wizard’s and User’s utterances are shown on separate tracks.

The ANVIL tracks “userUtts” and “wizardUtts” have the following properties:

Property	Description	Example(s)
transcription	Indicates the actual words spoken, as transcribed by the transcriptionist	not too expensive
transComments	Optional comments added by the transcriptionist.	he coughs at the end
turnHasFloor	Provides a link to the turn(s) (if any) in which this utterance has the floor	(ANVIL multi-link type)
turnNotFloor	Provides a link to the turn(s) (if any) in which this utterance <i>does not</i> have the floor	(ANVIL multi-link type)
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 7: Properties of the *states* ANVIL track in the Full-Audio Dialogues

“turnsFullAudio” track

The “turnsFullAudio” track expresses the “turn” information for the full-audio dialogues as described above.

The “turnsFullAudio” track is an independent ANVIL track – i.e., unlike the “turns” track, it is not an ANVIL span of another track. So that a researcher may extract all utterances covered in one turn, links are provided which connect the turn to all covered utterances.

The ANVIL track “turnsFullAudio” has the following properties:

Property	Description	Example(s)
who	Identity of the active participant in this turn.	Wizard, User
uttFloor	Provides link to the one or more utterance(s) that make up this turn.	(ANVIL multi-link type)
uttNotFloor	Provides link to the utterance(s) (if any) which fail to take over this turn.	(ANVIL multi-link type)

Table 8: Properties of the *turnsFullAudio* ANVIL track in the Full-Audio Dialogues

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	Incl in Full-Audio?	Incl in Sim-ASR?	Example(s)
FileName	The filename (minus the extension like .anvil or .mov) of this file.		Y	001-001-1
Date	The date on which this dialog was conducted.		Y	06-Feb
WizardID	The ID of the wizard in this dialogue. Wizard IDs range from 001 to 012 for simulated ASR channel dialogues; wizards 013 and 014 participated in full-audio dialogues.		Y	1, 2, ... 14
UserID	The ID of the user in this dialogue. User IDs range from 001 to 036 for simulated ASR channel dialogues, and from 037 to 042 for full-audio dialogues.		Y	1, 2, .. 42
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.		Y	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. <i>FullAudio</i> indicates this was a full-audio dialog.		Y	None, Low, Med, Hi, FullAudio
TaskID	The ID of the task used in this dialogue. There are a total of 24 tasks.		Y	1, 2, ... 24
UserOrder	The order of this dialogue as experienced by the user. Each user		Y	1, 2, 3, 4

	undertook 4 dialogues.			
WizardOrder	The order of this dialogue as experienced by the wizard. Each wizard undertook 12 dialogues -- 4 dialogues with 3 different users.		Y	1, 2, ... 12
WizardUser	The order of this user as experienced by this wizard. Each wizard interacted with 3 different users.		Y	1, 2, 3
TaskPrecision	A grading of precision of the answer provided by the user. See discussion of task completion, above.		Y	na, 1, 2, 3
TaskRecall	A grading of the recall of the answer provided by the user. See discussion of task completion, above.		Y	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").		Y	True, False
WizardLikert{1..7}	Wizard's response to Likert-scored question n on Form W-Q for this dialogue.		Y	1, 2, ... 7
UserLikert{1..7}	User's response to Likert-scored question n on Form S-Q for this dialogue.		Y	1, 2, ... 7

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 ASR-channel specification file, `woz-anvil-spec.xml`. The ANVIL file next references the associated audio (video) file `999-888-7.mov`.

In this example, the user says “hi i'm looking for a hotel for me and my partner”, which is misrecognized as “hi i'm looking for a hotel for me um my corner”. The wizard responds with “okay so what sort of hotel would you like to have.”

```

<?xml version="1.0" encoding="ISO-8859-1"?>
<annotation>
  <head>
    <specification src="woz-anvil-spec.xml" />
    <video src="999-888-7.mov" />
  </head>
  <body>
    <track name="states" type="primary">
      <el index="0" start="0" end="0.25">
        <attribute name="type">SILENCE</attribute>
        <attribute name="next">USER_TALKING</attribute>
        <attribute name="cmd">start</attribute>
        <attribute name="id">ep.typistears</attribute>
      </el>
      <el index="1" start="0.25" end="7">
        <attribute name="wavFile">user-082--2004-02-06-12-32-04.wav</attribute>
        <attribute name="type">USER_TALKING</attribute>
        <attribute name="next">TYPIST_TYPING</attribute>
        <attribute name="cmd">stop</attribute>
        <attribute name="id">ep.typistears</attribute>
      </el>
      <el index="2" start="7" end="10">
        <attribute name="type">TYPIST_TYPING</attribute>
        <attribute name="next">CONFUSER_CONFUSING</attribute>
        <attribute name="cmd">transcription</attribute>
        <attribute name="transcription">HI I'M LOOKING FOR A HOTEL FOR ME AND MY
          PARTNER</attribute>
        <attribute name="id">tm.typist</attribute>
      </el>
      <el index="3" start="10" end="11">
        <attribute name="confused">HI I'M LOOKING FOR A HOTEL FOR ME UM MY
          CORNER</attribute>
        <attribute name="type">CONFUSER_CONFUSING</attribute>
        <attribute name="next">SILENCE</attribute>
        <attribute name="cmd">confused</attribute>
        <attribute name="id">confuser</attribute>
      </el>
      <el index="4" start="11" end="12">
        <attribute name="type">SILENCE</attribute>
        <attribute name="next">WIZARD_TALKING</attribute>
        <attribute name="cmd">start</attribute>
        <attribute name="id">ep.userears</attribute>
      </el>
      <el index="5" start="12" end="19">
        <attribute name="id">ep.typistears</attribute>
        <attribute name="cmd">start</attribute>
        <attribute name="transcription">okay so what sort of hotel would you like
          to have</attribute>
        <attribute name="next">USER_TALKING</attribute>
        <attribute name="wavFile">wiz-072--2004-02-06-12-32-16.wav</attribute>
        <attribute name="type">WIZARD_TALKING</attribute>
      </el>
      ...
    </track>
  </body>
</annotation>

```

```

<track name="turns" type="span" ref="states">
  <el index="0" start="1" end="3">
    <attribute name="groundingAct02">Request</attribute>
    <attribute name="groundingAct01">GreetingFarewell</attribute>
    <attribute name="prevUserTurnUnderstanding">NA-User</attribute>
    <attribute name="who">User</attribute>
  </el>
  <el index="1" start="5" end="5">
    <attribute name="groundingAct02">Request</attribute>
    <attribute name="groundingAct01">ExplAck</attribute>
    <attribute name="prevUserTurnUnderstanding">Full</attribute>
    <attribute name="who">Wizard</attribute>
  </el>
  ...
</track>
<set name="dialogProps">
  <el index="0">
    <attribute name="FileName">999-888-7</attribute>
    <attribute name="Date">06-Feb</attribute>
    <attribute name="WizardID">26</attribute>
    <attribute name="UserID">96</attribute>
    <attribute name="ErrorRate">Med</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." *Proc. ICSLP*, October 2004, Jeju, South Korea.
- [2] Jason D. Williams and Steve Young. (2004). "Characterizing Task-Oriented Dialog using a Simulated ASR Channel." *Proc. ICSLP*, October 2004, Jeju, South Korea.
- [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] Gabriel Skantze, 2003. "Exploring human error handling strategies: implications for spoken dialogue systems." *Error Handling in Spoken Dialogue Systems*, ISCA Tutorial and Research Workshop, Château d'Oex, Vaud, Switzerland, August 28-31, 2003, pp 71-76.
- [14] Christine Duran, John Aberdeen, Laurie Damianos and Lynette Hirschman, 2001. *Comparing several aspects of human-computer and human-human dialogues*. In *Proceedings of the 2nd SIGDIAL Workshop on Discourse and Dialogue*, Aalborg, 1-2 September 2001: 48-57.
- [15] Stockle et al., 2000. *Dialog Act Modeling for Automatic Tagging and Recognition of Conversational Speech*.
- [16] Taylor et al., 1998. Intonation and dialog context as constraints for speech recognition. *Language and Speech* 41(3-4):489-508.
- [17] Jekat et al. 1995. *Dialogue acts in Verbmobil*. *Verbmobil-Report* 65.
- [18] Nagata, Masaaki. 1992. Using Pragmatics to rule out recognition errors in cooperative task-oriented dialogues. *Proc. Intl. Conf on Spoken Language Processing*. Vol 1, pp 647-650, Banff, Canada, Oct 1992.
- [19] David R. Traum, "A Computational Theory of Grounding in Natural Language Conversation", TR 545 and Ph.D. Thesis, Computer Science Dept., U. Rochester, December 1994.

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 *2*, this indicates the document was used for the direct audio channel experiments.

Experiments	Who	Name	Description	Page
All experiments	Subject	<i>Form S-T</i>	Tasks	25
		<i>Form S-M</i>	Subject Map	49
	Wizard	<i>Form W-T</i>	Task information	50
		<i>Form W-M</i>	Wizard Map	56
Simulated ASR Channel Experiments	Subject	<i>Form S-R</i>	Role description	57
		<i>Form S-Q</i>	Questionnaire	58
	Wizard	<i>Form W-R</i>	Wizard	59
		<i>Form W-Q</i>	Questionnaire	60
	Typist	<i>Form T-R</i>	Role description	61
Direct Audio Channel Experiments	Subject	<i>Form S-R-2</i>	Role description	62
		<i>Form S-Q-2</i>	Questionnaire	63
	Wizard	<i>Form W-R-2</i>	Role description	64
		<i>Form W-Q-2</i>	Questionnaire	65

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 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:

- En suite rooms
- Quiet rooms
- 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 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 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 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:

- (1) A list of hotels and their price ranges
- (2) 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 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.

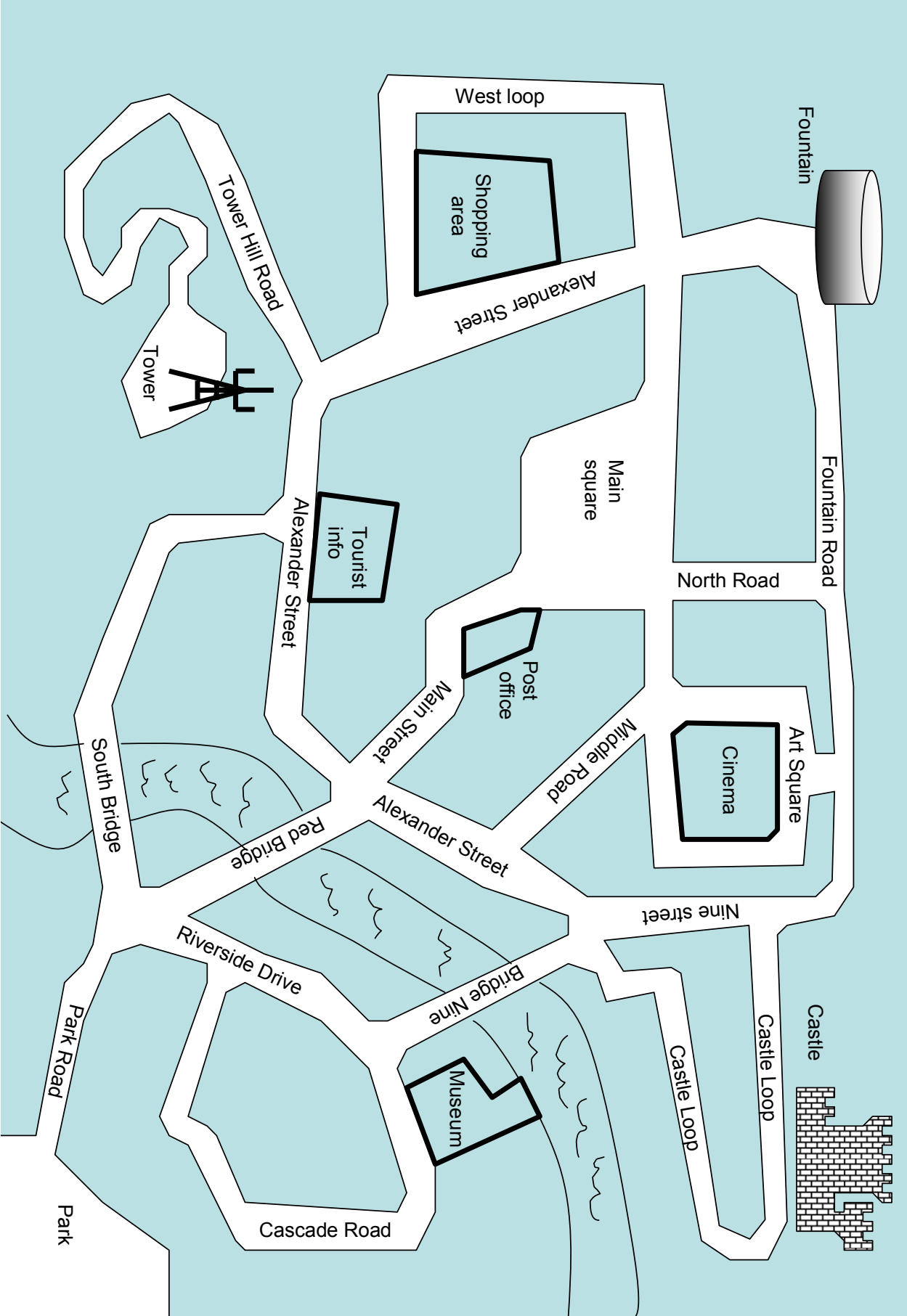
You'll be walking, so try to minimize the distances between the bars.

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



TOWN INFORMATION
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

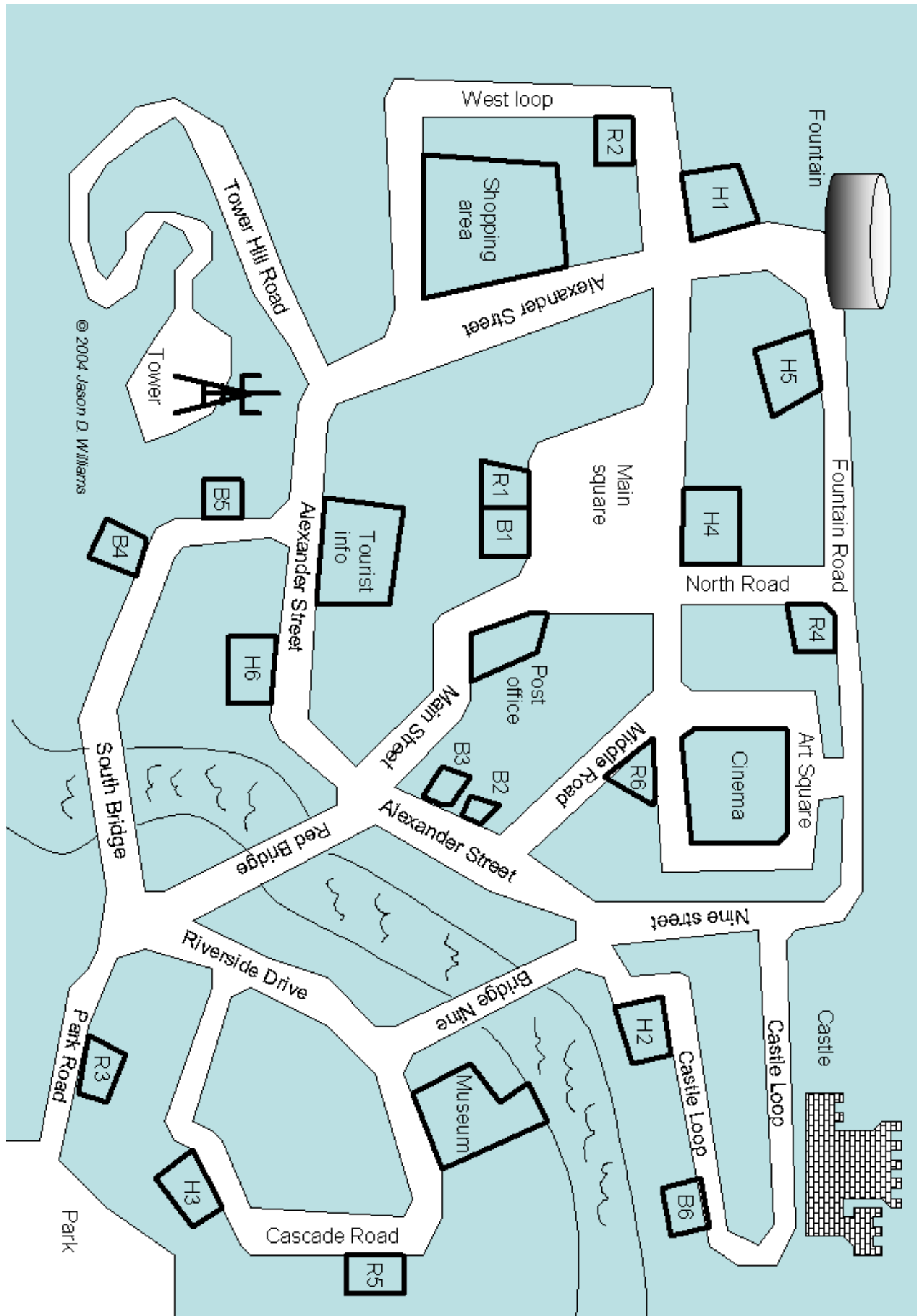
Shown in yellow on the map

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



© 2004 Jason D. Williams

ROLE DESCRIPTION

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)
-----------------------------	-----------------	-----------------------------	--------------------------------------	-----------------------	--------------	-----------------------

ROLE DESCRIPTION

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:

- 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 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.

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)
-----------------------------	-----------------	-----------------------------	--------------------------------------	-----------------------	--------------	-----------------------

ROLE DESCRIPTION

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.

ROLE DESCRIPTION

In this study, you will be asked to complete several tasks, in cooperation with another person, who is also a test subject. The other subject is in a different room, and you will interact with the other subject using headphones and a microphone.

In this study, you will be playing the part of a tourist, and the other subject will be playing the part of a tour guide. 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-2

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 was clearly understood by the other subject.*

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 headphones and microphone.*

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 headphones and microphone 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 task.*

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

ROLE DESCRIPTION

In this study, you will be asked to assist other subjects to complete several tasks. The other subject will be in a different room, and you will interact with the other subject using a pair of headphones and a microphone. When you speak, you'll also be able to hear yourself, similar to a land-line telephone.

In this study, you are playing the part of a tour guide, and the other subject is playing the part of a tourist who 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:

- 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 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.

QUESTIONNAIRE

Form W-Q-2

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 could clearly understand the other subject.*

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 headphones & microphone.*

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 headphones and microphone 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 task.*

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

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.

- o i thought he -- i thought he was there
- o 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:

- o 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.:

- o (())

- [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:

- o do you dress like a *schlump yet
- o 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].
 - o [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.