UNIVERSITY OF GOTHENBURG

DEPARTMENT OF PHILOSOPHY,
LINGUISTICS AND THEORY OF SCIENCE

GROUNDING OF NAMES IN DIRECTORY ENQUIRIES DIALOGUE

A corpus study of listener feedback behaviour

Anastasia Bondarenko

Master's Thesis:	30 credits
Programme:	Master's Programme in Language Technology
Semester and year:	Spring, 2019
Supervisors	Staffan Larsson, Christine Howes, David Hjelm
Examiner	Lars Borin

Abstract

This paper presents a new corpus of dialogues in the domain of directory enquiries. We describe its collection and annotation process and then analyse feedback strategies employed by the dialogue participants focusing mainly on the grounding instances in the context of transmission of names. We discuss our findings in regards to their implementation in dialogue systems as well as in comparison to previous corpus studies of feedback. Finally, we present a preliminary formalisation of the grounding process of names, using a finite-state approach to modelling grounding in dialogue proposed by Traum (1994).

Contents

1 Introduction 1

2 Background 2

2.1 Grounding in dialogue 2

2.2 Corpus studies of feedback in dialogue 5

2.3 Feedback in the domain of telephone directory enquiries 8

3 Directory Enquiries Corpus 9

3.1 Data collection 9

3.2 Transcription and annotation 10

4 Preliminary analysis of the corpus 12

4.1 Quantitative results 12

4.2 Qualitative results 16

4.3 Finite-state model of grounding names 23

5 Conclusions and future work 28

References 30

Appendices 33

Appendix 1 33

Appendix 2 34

Appendix 3 35

1 Introduction

Dialogue has been described as a collective activity whereby the participants engage in a joint effort of moving the conversation forward (Clark & Schaefer, 1987; Cahn & Brennan, 1999). This is mainly achieved as a result of the participants' reliance on continuous use of specific utterances that can be classified as listener feedback whose purpose is to indicate some kind of evidence of understanding or misunderstanding between the dialogue partners. In other words, these utterances indicate whether a conversational proposition has been grounded or not.

The motivation behind this master's project comes from the need to consider communicative grounding strategies, which are characteristic of human-human dialogue, in the context of spoken dialogue systems, for the purpose of providing a basis for selecting appropriate strategies to be implemented in such systems. More specifically, we focus on listener feedback in the domain of telephone directory enquiries and argue that this particular restricted domain can provide a good basis for studying feedback behaviours in human-human dialogue as well as a good test case for implementing such behaviours in dialogue systems.

Thus, we have collected and published a new corpus of dialogues belonging to the domain of directory enquiries in an effort of providing a collection of relevant data to facilitate the investigation of feedback strategies used by human participants thus aiding in the development of spoken dialogue systems that deal with contexts where a successful transfer of accurate information between the user and the system is crucial. We further annotate and explore the collected data focusing mainly on the process of grounding of names as we argue that there is a greater chance of miscommunication when it comes to the transmission of such content between participants as opposed to communicating number sequences. In this regard, we provide a preliminary analysis of feedback utterances found in the corpus including acknowledgements and clarification requests, as well as an investigation of specific ways they are utilised by the dialogue participants. Additionally, we discuss our findings as applied to dialogue systems implementations and compare them to previous corpus studies of feedback. Finally, we provide a preliminary formalisation of the process of grounding of names, thus, proposing a finite-state model adopting Traum's approach to modelling grounding in dialogue (Traum, 1994).

Section 2 provides a survey of previous work that serves as the background for this thesis, including the theory of conversational grounding, corpus studies of feedback in dialogue, as well as the context for studying feedback specifically in the domain of telephone directory enquiries. Section 3 presents a new directory enquiries corpus and describes its collection, transcription and annotation. Section 4 reports the results of a preliminary analysis of the collected corpus, both quantitative and qualitative, and provides a discussion on its findings. Additionally, this section presents a tentative finite-state model of grounding of names based on the analysed data. Finally, section 5 summarises the main points of the thesis and suggests potential directions for future work.

2 Background

2.1 Grounding in dialogue

Effective communication requires collaboration between all participants. In this regard, Clark & Schaefer (1987) discuss the collaborative effort that needs to be demonstrated by both the speaker and the addressees during the communicative process. They further argue that this effort is based on the concept of *common ground*, which is often defined as a set of *mutual beliefs* or *shared information* (Traum, 1994) that the speaker and the addressees continually update in a process referred to as *grounding*.

Grounding can be reached through a continuous process of specification and subsequent acceptance of a coherent piece of information by a speaker, according to Clark & Schaefer (1987), who describe this collaborative process in terms of the participants' *contributions* to the conversation. They specify that each contribution consists of a *presentation phase*, which initiates the contribution, and an *acceptance phase*, which is supposed to result in mutual acceptance of the original or revised contribution and is considered to have taken place once the contribution meets *the grounding criteria* of each participant (Clark & Schaefer, 1987).

> **Grounding criterion:** The contributor and the partners mutually believe that the partners have understood what the contributor meant to a criterion sufficient for current purposes. (Clark & Schaefer, 1989, p. 262)

In other words, each participant has to decide whether their partner has demonstrated sufficient evidence that they have accepted the presented contribution for it to be added to the common ground. The grounding criteria are context-dependent and domain-dependent. In some types of dialogue, e.g. air traffic control, they can even be formally and legally stipulated (Cahn & Brennan, 1999).

In a conversation between speaker A and speaker B, B initiates the acceptance phase by presenting some kind of evidence of understanding of A's contribution. (Clark & Schaefer, 1989, p. 267) define the five types of such evidence and grade them from weakest to strongest as follows:

1. **Continued attention:** B shows that he is continuing to attend and therefore remains satisfied with A's presentation
2. **Initiation of relevant next contribution:** B starts in on the next contribution that would be relevant at a level as high as the current one
3. **Acknowledgement:** B nods or says "uh huh", "yeah", or the like
4. **Demonstration:** B demonstrates all or part of what he has understood A to mean
5. **Display:** B displays verbatim all or part of A's presentation

The grounding criterion, in this regard, incorporates the notion of the strength of evidence, which, in turn, depends on the type of presentation it follows. That is to say that the domain and the context play a major role in what degree of evidence speaker B deems appropriate to demonstrate at any given point and whether speaker A would regard this evidence as sufficient (Clark & Schaefer, 1989). A person making a doctor's appointment on the phone, for example, might repeat the suggested time and date back to the receptionist, i.e. display A's presentation verbatim, while a person listening to their friend recount the events of the past week might simply maintain eye contact and not interrupt their partner, i.e. show continued attention.

It is not always possible to reach complete understanding in a conversation, however, in which case the addressee might have to somehow indicate this lack of understanding. The problems in understanding may arise on several different levels. In this respect, (Clark & Schaefer, 1989, p. 268) define *u'* as part of or whole of A's presentation and describe the four states of understanding as follows:

1. **State 0:** B didn't notice that A uttered some *u'*
2. **State 1:** B noticed that A uttered some *u'* (but wasn't in state 2)
3. **State 2:** B correctly heard *u'* (but wasn't in state 3)
4. **State 3:** B understood what A meant by *u'*

In a situation where speaker B has not been able to reach the final state, state 3, A's proposition cannot be accepted. For it to happen, B's "general strategy is to initiate a side sequence to get A to help him reach state 3" (Clark & Schaefer, 1987, p. 22). This might manifest in a number of different responses on B's part from conventional phrases like "Excuse me?" or "Pardon?" if B is in state 1 (or state 2), to such requests as "And?" or "What do you mean?" if B is in state 2, for example. Each of these requests is a contribution in itself, according to the contribution model, and is considered to be subordinate to A's initial contribution.

The types of utterances that seek to elicit confirmation or correction from a conversational partner are usually referred to as *other-repairs*. According to the principle of least collaborative effort, however, *self-repairs* are more typical of conversations (Clark & Schaefer, 1989; Schegloff et al., 1977). This means that speaker A will construct their contribution in such a way as to reduce the combined effort spent on presenting and accepting that contribution. In practice it means that the speaker presenting the contribution will try to correct any errors or potential misunderstandings themselves rather than expect their partner to ask for clarifications.

The contribution model is modified and further extended by Cahn & Brennan (1999) to make it more appropriate for modelling human-computer dialogue. Adopting an approach which emphasises that each dialogue state representation in the model reflects the point of view of one of the dialogue participants rather than taking into account all perspectives, they propose to model dialogue in terms of participants' *exchanges*. An exchange, in this context, is a minimal dialogue unit consisting of two contributions or meaningful pairs of utterances. Based on the idea of *adjacency pairs* introduced by Schegloff & Sacks (1973) this model seeks to capture the two-part structure and the collaborative nature of conversational tasks.

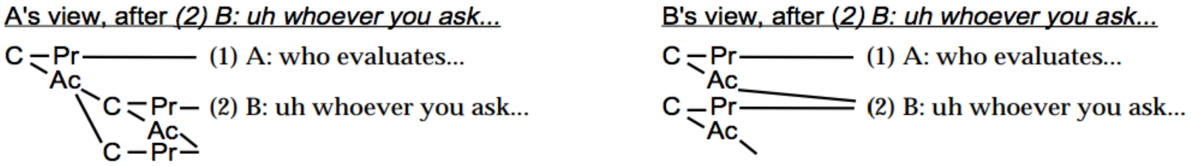

A's view, after (2) B: uh whoever you ask...
C
Pr
(1) A: who evaluates...
Ac
C
Pr
(2) B: uh whoever you ask...
Ac
C
Pr
B's view, after (2) B: uh whoever you ask...
C
Pr
(1) A: who evaluates...
Ac
C
Pr
(2) B: uh whoever you ask...
Ac

Figure 1: Divergent models from the point of view of each conversation participant (reproduced from Cahn & Brennan, 1999, p. 3)

Cahn & Brennan (1999) claim that the idea to keep divergent representations of the dialogue from the point of view of each participants (Figure 1) stems from the need to eliminate possible confusion regarding whose perspective is being reflected in the graph and whether it is shared. This primarily concerns dialogues that contain repairs, where maintaining a coherent node structure to keep up with everyone's point of view quickly becomes problematic.

Similar discussion of the contribution model appears in Traum (1994) where one of the deficiencies mentioned is the fact that the model lacks clarity in terms of the distinction between the phases. Since repairs in general and other-initiated self-repairs in particular usually result in different kinds of clarification subdialogues within a conversation it is usually hard to tell whether such utterances belong to the presentation phase or the acceptance phase. As such, in order to make any judgement as to where a particular utterance fits in the model one has to consider the conversation as a whole.

Furthermore, he argues that the presentation-acceptance structure of the contribution model as it is is not sufficient to make decisions about any possible next utterances at any given point during a conversation. To illustrate this he uses the following two examples (reproduced from Traum, 1994, p. 32):

(1)

1	A	Move the boxcar to Corning
2	A	and load it with oranges
3	B	ok

(2)

1	A	Move the boxcar to Corning
2	B	ok
3	A	and load it with oranges
4	B	ok

According to the contribution model the exchange in Example 1 represents a single contribution and the first two utterances are part of the presentation phase, as opposed to Example 2 where there seem to be two separate contributions and both utterances by speaker A are separate presentations. This potentially means that the only way to assert that a presentation has ended is when the subsequent utterance has already been put forward (Traum, 1994).

Consequently, Traum (1994) substitutes the concept of contributions with *discourse units (DUs)*, i.e. units of dialogue at which grounding takes place. These discourse units, in turn, instead of phases are divided into actions (or acts) represented by individual utterances:

- **Initiate:** opening utterance
- **Continue:** subsequent utterances which add new information
- **Acknowledgement (Ack):** utterance that claims understanding of a preceding utterance
- **Cancel:** utterance that makes previous propositions ungroundable
- **Repair:** utterance that changes the material under consideration
- **Repair request (ReqRepair)**: the responder requests a clarification/correction from the initiator
- **Acknowledgement request (ReqAck)**: the initiator requests an evidence of understanding from the responder

In the resulting finite-state model state S is an uninitiated discourse unit, state F is a grounded discourse unit and state D is an ungrounded or ungroundable discourse unit. The other four states shown in Table 1 require a number of additional utterance acts to be considered grounded. For example, in state 1 what is needed is for the responder to provide an acknowledgement of the initiator's utterance, state 2 is a case of other-initiated self-repair where the initiator needs to clarify or amend their previous proposition, and so on.

Modeling grounding in dialogue in such a way presupposes that each grounding instance is achieved after a specific sequence of actions has been performed by the two participants. In an event where an action is performed that falls outside of the model's scope it implies that in order for the conversation to proceed

State	Entering act	Preferred exiting act
S	-	Initiate^{I}
1	Initiate^{I}	Ack^{R}
2	ReqRepair^{R}	Repair^{I}
3	Repair^{R}	Ack^{I}
4	ReqRepair^{I}	Repair^{R}
F	Ack$^{\{I,R\}}$	Initiate$^{\{I,R\}}$ (next DU)
D	Cancel$^{\{I,R\}}$	Initiate$^{\{I,R\}}$ (next DU)

Table 1: Discourse unit states where I is the initiator and R is the responder (reproduced from Traum (1994) p. 42)

there needs to be a repair of some kind or that a completely new discourse unit should be initiated (Traum, 1994).

2.2 Corpus studies of feedback in dialogue

Taking into account the collaborative theory of human-human communication it is important to reiterate that one of the characteristic features of dialogue is that it is continuously co-constructed by all participants. This is achieved by employing specific communicative grounding strategies to move the conversation forward. Listeners provide frequent *feedback* to indicate whether they have been able to ground the information provided by the speakers. In general such feedback comes in the form of relevant next contributions, or *backchannels* (e.g. “mmhm”, “okay”, Example 3, lines 6 and 8).[1]

Another type of response is a *clarification request*, which signals misunderstanding and uncertainty of some sort or a lack of perception or coordination and usually implies a need for repair (Example 3, lines 11, 17, 20, 24, 26, 31).

(3) DEC07:1–32

1	Caller	hello
2	Operator	hello
3	Caller	hello
4	Operator	how may i help you?
5	Caller	oh hi i’m uh looking for some phone numbers
6	Operator	yes
7	Caller	er here in london
8	Operator	yeah
9	Caller	and the first
10		one is rowans tenpin bowl
11	Operator	can you repeat that for me?
12	Caller	rowans tenpin bowl
13		so it’s rowan
14		R O W A N S
15	Operator	yes
16	Caller	tenpin
17	Operator	tenpin?
18	Caller	yeah
19	Operator	the number ten

[1] Dialogue examples labeled with DEC are all taken from our Directory Enquiries Corpus presented in this thesis

20		and pin?
21	Caller	yes
22		yes
23	Operator	tenpin
24		road?
25	Caller	bowl
26	Operator	th- like the bird?
27	Caller	uh like bowling
28	Operator	uh bowling
29	Caller	bowl
30	Operator	yes
31		the thing you eat from right?
32		okay here we go

Although it seems somewhat difficult to compare corpus studies of feedback due to the associated terms such as backchannel, acknowledgement, clarification and others being used rather inconsistently in the literature and even deemed quite problematic by some (Fujimoto, 2007), there are a number of related studies that bear mentioning. One of the earliest studies of backchannels is found in Duncan (1972, 1974), where an analysis of transcribed face-to-face conversations between pairs of people is provided. The two conversations used in this study are video recordings of a preliminary interview between a therapist and his new client, and a work-related discussion between two therapists. Backchannels in the context ot this analysis are defined as a way for the listener to provide the speaker with useful information while the speaker's turn is ongoing. The claim here is that backchannel utterances do not constitute a separate turn. The types of backchannel signals according to this study are "mhm" signals, sentence completions, requests for clarification, brief restatements, as well as head nods and shakes. The study finds that in 885 "units" (roughly corresponding to utterances) there are a total of 71 instances of feedback (8%).

Cerrato (2002) analysed real dialogues between travel agents and customers taken from The Gothenburg University Spoken Language Corpus (Allwood, 1999) and found that "feedback expressions" of the customers comprised more than 50% of all individual turns. In this study feedback expressions are expressions showing "continuation of contact", acknowledgements (among which are, for example, short expressions like "ja", repetitions or reformulations of part or the whole of the previous utterance, gestures and facial expressions), as well as expressions of agreement/disagreement.

Another corpus study that covers aspects of feedback is by Fernández (2006) who is focusing on non-sentential utterances (NSUs) in the dialogues taken from the British National Corpus (BNC) (Burnard, 2000). The dialogues used in this work are selected at random and belong to a number of different domains including meetings, tutorials, interviews, medical consultations, as well as free conversations. This study includes annotations of such classes as acknowledgements (5% of all utterances), and clarification ellipsis (1%). However, the numbers reported here might be an underestimate as for the total number of feedback utterances since the author deliberately excludes cases of utterances produced in overlap (Rühlemann, 2007) as well as sentential cases of clarification requests (e.g. "what do you mean?").

One more study based on BNC dialogues is specifically focused on clarification requests (Purver, 2004) and is quite detailed in the way it categorises them into different types. 90% of the data used in this study is "general non-context-governed dialogue recorded by subjects during their daily lives" with the rest being conversations belonging to "various domains" (Purver, 2004, p. 57). All CRs here are classified according to their form: non-reprise clarifications ("what did you say?"), reprise sentences ("we are going on wednesday" - "you are going on wednesday?"), wh-substituted reprise sentences ("you are going when?"), reprise sluices ("when?"), reprise fragments ("wednesday?"), reprise gaps ("you are going on?"), gap fillers ("i got the"

- "flowers?"), conventional ("pardon?"). The total percentage of CRs in the dialogues in this study is just under 3% with conventional and reprise fragment forms being the most common ones and each comprising about 30% of the total number of CRs.

Colman & Healey (2011) compared patterns of repair in the BNC dialogues and the HCRC Map Task corpus (Anderson et al., 1991) and found that CRs were more frequent in the latter with the overall frequency of "repair events" being twice as much in the more processing demanding task-oriented domain. The frequency of CRs was also found to vary depending on the role of the speaker in the task. Namely, route followers produced significantly more clarification requests than route givers.

Similarly, Rieser & Moore (2005) conducted a study of CRs in task-oriented conversations and compared the results to previous studies. Their findings support the claim that CRs are more characteristic of task-oriented dialogue. Beyond that, they report that most CRs are partial in form and almost all of them directly follow the problematic utterance. It is also noted that the CR-initiators tend to ask their partners to confirm a hypothesis about a perceived utterance instead of asking them to repeat the utterance in question.

Moreover, feedback has been studied in the context of human-computer interaction. Ward & Heeman (2000) conducted an experiment exploring the user's willingness for using unprompted acknowledgements and repetitions in a human-computer dialogue which demonstrated that about a half of all participants utilised these two strategies at least occasionally, and that around 29% of them used acknowledgements more often than specific system commands.

Bell & Gustafson (2000) investigated feedback strategies in a Swedish human-computer interaction corpus and discovered that positive ("yes", "good") and negative ("no", "well") feedback indicated understanding and misunderstanding in the dialogues respectively, and suggested that positive feedback could be used as a way for the system to learn the user's preferences, while negative feedback could prevent serious errors from occurring. Additionally, in their study they distinguished between explicit ("that's great") and implicit ("mhm") feedback and ascertained that two thirds of all feedback utterances in the corpus belonged to the explicit group with one third of them belonging to the implicit one. Skantze et al. (2006) studied task-oriented human-computer dialogue and among other things reported that brief acknowledgements such as "yes" and "mm" were the most common type of feedback provided by human users following a system utterance.

In view of this, there has been an increased focus on the notion of incrementality in regards to processing speech within dialogue systems. Incremental models of dialogue, as opposed to turn-taking ones, are able to process user input and generate a system response not only while the user is speaking but also while they are listening. This has been found to result in dialogue that is more "human-like" and generally more preferred by users than the one traditional models are capable of. Additionally, incremental processing allows for faster error correction which improves the overall user experience (Skantze & Schlangen, 2009). Hjalmarsson & Edlund (2008) found that between a dialogue system that demonstrates such human-like conversational behaviours as grounding, hesitations and fragmental utterances and a system with more constrained utterance generation, the former was perceived as more "intelligent and polite" with no significant effect on the overall efficiency of the conversation.

Skantze & Schlangen (2009) presented an incremental micro-domain (number dictation) dialogue system that is capable of producing rapid feedback while the user's turn is ongoing as well as reacting to the user's feedback during its turn respectively. Continuing this work Buß et al. (2010) further extend the system's capabilities to accommodate grounding in semantically more complex domains where understanding at state 3 rather than state 2 (as per previously mentioned understanding levels introduced by Clark & Schaefer, 1989), which was sufficient in the case of number sequences dictation, is required. Specifically this extension supports the system's ability to produce "overlapping non-linguistic" feedback (e.g. "erm") to prompt the user for a clarification or reformulation. Other models that incorporate incremental grounding in dialogue systems have been proposed, including the ones focusing on listener feedback in multi-party conversations (Wang et al., 2011) and overlapping feedback behaviour (Visser et al., 2014; Khouzaimi et al., 2014).

2.3 Feedback in the domain of telephone directory enquiries

The focus of this thesis is on listener feedback in a very restricted domain of telephone directory enquiries (DE). DE systems have been an important application case for dialogue systems for several decades now (Chang, 2007). During this time it has become standard practice to employ specific automated processes as part of such systems, which led to their commercial success and an increased worldwide availability (van Heerden et al., 2014). However, there still remain some challenges in making such systems perform in a fully-automated way while at the same time being a reliable and convenient service for its users.

One of the problematic aspects that arise in the context of this particular domain is the system's ability to understand names that are not present in an existing lexicon. This is even more crucial for conversations on the phone where a noisy environment and the inability for the participants to see each other can lead to failures in understanding.

In addition to the implementation-related goals directory enquiries is also a particularly good domain for studying feedback due to its task-oriented nature. This means that feedback utterances in such conversations are more frequent than in less goal-driven domains as a result of the participants' commitment to successful information transfer (Colman & Healey, 2011). Additionally, the increased frequency of verbal feedback in this domain can be attributed to the impossibility of relying on non-verbal feedback as part of the interactions on the phone where the interlocutors cannot see each other (Boyle et al., 1994). Moreover, unlike other data sources for studying dialogue, such as the aforementioned Map Task for example (Anderson et al., 1991), the task of a DE call is less asymmetric because both participants have a chance to act as an "information giver" and an "information receiver" at different stages of the conversation. Specifically, the caller has to communicate the name of the query to the operator and the operator has to relay the enquired phone number back to the caller with both of them needing to ground the received information.

That said, it is important to note that there is an apparent paucity of data available for investigating dialogues in this specific domain. Although there are some corpora that incorporate DE calls as a type of dialogue, such as the Estonian Dialogue Corpus (Koit, 2012), or have annotations that include grounding acts of personal names, such as the Loqui Human-Human Dialogue Corpus (Passonneau & Sachar, 2014), there is a distinct lack of English-language dialogue data that can provide insight into feedback behaviour in the context of this particular task. Furthermore, what motivates us to focus our attention on this domain is the fact that such data, as opposed to the corpora mentioned before, does not require anonymisation since all of the relevant information (such as business names and phone numbers) is publicly available.

3 Directory Enquiries Corpus

3.1 Data collection

The data was collected with the help of 14 volunteers who were paired up for each recording session. Eight of the volunteers were male and six of them were female. As can be seen in Table 3, the participants were native speakers of a number of different languages and had various levels of English proficiency.

First language(s)	Number
Swedish	6
Hindi	2
English	1
Hungarian	1
Icelandic	1
Norwegian	1
Spanish, Catalan	1
Spanish	1

Table 3: Participants' first languages

The pairs of participants were given instructions on the recording setup (see Appendix 1 and 2). Each of them were to take turns playing the roles of a directory service operator and caller. Each caller had been provided with a list of businesses located in London. The lists had 3 business names per recording location, the description of their types, as well as their addresses. Since the majority of the participants were non-native English speakers the names of the businesses had been selected in such a way that every list had a combination of names with varying degrees of complexity (e.g. "The Good Earth" vs. "Hawksmoor Spitalfields"). This was done in an attempt to simulate a situation where the operator has to deal with a name that is potentially out of their lexicon.

In this regard, it is important to emphasise that the collection of data was based on a simulation of a directory enquiries interaction rather than a real one, which means that the collected dialogues might not constitute a representation of general communicative behaviours exhibited by professional operators. However, since both participants in the simulation act based on their understanding of the provided instructions and are both goal-driven, the data does provide valuable insight into the participants' behaviours demonstrated within the specific tasks of name and number sequence grounding, which was the main goal of the data collection and subsequent analysis.

The caller's task was to call up the operator and ask for the phone numbers of the businesses on their list. The operator's task, in turn, was to provide the caller with the necessary information using the online Phone Book service (thephonebook.bt.com). Each caller had to make one call at an inside location with a low to moderate noise level and another one at an outside location with a moderate to high noise level. Operators were receiving the calls at the studio. Each speaker was recorded with microphones placed at each recording location. The recording sessions resulted in 4 dialogues per pair with the shortest dialogue's duration being 2 minutes 31 seconds and the longest one being 10 minutes 46 seconds. Thus, 28 dialogues were collected in total.

Each volunteer was asked to sign a consent form making it possible for the collected data including the audio recordings, transcriptions, annotations and metadata to be made freely available on the Open Science Framework (osf.io/2vjkh; Bondarenko et al., 2019).

3.2 Transcription and annotation

The audio recordings were transcribed with the help of the ELAN annotation tool (Brugman & Russel, 2004). It is worth noting that in order to facilitate annotation the utterances produced with rising intonation include a question mark in the transcripts (e.g. "sorry?").

In the next step all of the transcripts were manually annotated. Two dialogues (281 utterances) were annotated by two coders to ensure inter-rater reliability. Cohen's kappa tests showed good agreement for the main tags: `turn-type (ack/CR/C)` $\kappa = 0.635$; `AckType` $\kappa = 0.625$; `CRType` $\kappa = 0.689$.

The overview of the main annotation tags is presented in Table 4. Our focus was on three main types of utterances that we tagged as ***acknowledgments (Ack)***, ***clarification requests (CR)*** and ***clarifications (C)***.

Tag	Value	Explanation
`acknowledgement (Ack)`	y/n	For all utterances: does this utterance contain a backchannel (e.g. "yeah", "mmhm", "right") or a repeated word or phrase acknowledging the proposition or speech act of a previous utterance? (Note that this category does not include direct answers to yes/no questions)
`clarification request (CR)`	y/n	For all utterances: does this utterance contain a clarification request, indicating misunderstanding of the proposition or speech act of a previous utterance?
`clarification (C)`	y/n	For utterances following a clarification request: does this utterance contain a response to a clarification request, clarifying the proposition or speech act of a previous utterance?

Table 4: Main annotation tags

Feedback utterances were further annotated into sub-types.[2] For acknowledgements these were:

- ***Continuer:*** acknowledgement/backchannel words like "okay", "yeah", "yes", "mmhm". This category does not include direct answers to questions, but rather those utterances that demonstrate that the preceding utterance has been understood/accepted
- ***Verbatim:*** verbatim repetitions of previous utterances or their parts ("seven two" – "seven two"; "N E Y S" – "N E Y S")
- ***Paraphrase:*** paraphrased repetitions of previous utterances or their parts. In essence these acknowledgements are attempts at verbatim repetitions that have not been successful rather than intentional paraphrases ("hawksmoor spitalfields" – "hawkswore spitalfields"; "seven zero six two" – "seven two six two"; "C U R" – "C U C")[3]
- ***Confirm:*** confirmation phrases like "correct", "exactly", "that's correct"

[2] Inter-rater reliability has not been established for this part of the annotation which constitutes an important direction for future work

[3] Based on the later discussions of the classification it was decided that the term ***misrendering*** might better reflect the intended meaning of the sub-category and is suggested for future improvements of the annotation

- ***Appreciate:*** appreciative response to a previous utterance (“great”, “good”, “perfect”)

For clarification requests these were:

- ***General request:*** Speaker 2 indicates a lack of perception/understanding of Speaker 1’s previous utterance (“sorry?”, “sorry, I didn’t get that”, “excuse me, what?”)
- ***Repeat request:*** Speaker 2 asks Speaker 1 to repeat their previous utterance (“come again?”, “can you repeat that for me?”, “could you start from the top?”)
- ***Confirmation request:*** Speaker 2 asks Speaker 1 to provide a confirmation (“the name of the place was A S A K U S A, right?”, “yeah?”, “three?”)
- ***Spelling request:*** Speaker 2 asks Speaker 1 to spell out the name of the queried business or its address (“could you spell that for me please?”, “is that a W?”)

It is important to recognise that in the classification we have presented above the categories for acknowledgements may conflate form and function, whilst those for clarification requests do not consider the form. This might mean that in our analysis we miss some important information regarding the differences and similarities between these two types of feedback utterances (Howes et al., 2019). This presents a potentially fruitful direction for future work.

Additionally, the annotations contain the information about a potential overlap between the speakers’ utterances. The ***PreviousTurnEndComplete*** tag, which is a version of the *end-complete* tag from Purver et al. (2009) applied to the utterances preceding the one in question, indicates the turns that occur before the previous turn can be considered complete. We also use the ***Continues*** tag from Purver et al. (2009) in the annotation. This tag makes it possible to observe which turns consist of multiple sub-utterances including the ones produced by the same speaker, as well as those attributed to the other speaker’s backchannels, pauses and other interruptions in speech.

Since the main purpose of the data collection was to investigate the domain of telephone directory enquiries the utterances were also labeled according to their content: namely, whether they include any information about the names, addresses and phone numbers of businesses. Each utterance labeled with any of these was then labeled according to the form such information was conveyed in:

- ***Word (part)***: speaker mentions the name of a business or its address in full or in part
- ***Spelling installment (part)***: speaker provides a spelling for the name or the address of a business in full or in part, usually in installments of a specific number of letters
- ***Number dictation installment (part)***: speaker dictates a phone number in full or in part, usually in installments of a specific number of digits

Similarly to the previous speaker and next speaker labels each utterance was further annotated with the content and form labels of the previous and the next utterance (***PreviousWord/spelling/dictation***, ***PreviousContent***, ***NextWord/spelling/dictation***, ***NextContent tags***). Figure 3 in Appendix 3 provides an example of how each utterance in the corpus was annotated.

4 Preliminary analysis of the corpus

4.1 Quantitative results

In the 28 collected dialogues there are a total of 4165 utterances produced over 3002 speaker turns. In our annotation and analysis[4] a turn is comprised of multiple consecutive utterances produced by the same speaker with no intervening utterances from the other conversation participant. Both the shortest dialogue and the longest one were recorded in the noisier of the two locations. The former consists of 64 utterances, or 48 turns, while the latter consists of 246 utterances, or 190 turns.

Overall, feedback utterances constitute 37% of all utterances in the corpus (52% of all turns). Furthermore, of all the utterances in the corpus 1285 are acknowledgements (31% of utterances and 43% of turns) and 277 are clarification requests (7% of utterances and 9% of turns). This is higher than reported in most previous studies (Duncan, 1972, 1974; Purver, 2004; Fernández, 2006) and might be domain and task-specific.

The percentage of feedback instances found in Cerrato (2002), however, is higher, which might be due to the fact that feedback in that study includes non-verbal signals, such as gestures and facial expressions, as well as utterances expressing agreement and disagreement, which was not the focus of our work. Additionally, that particular study, unlike ours, does not consider clarification requests as a type of feedback making it even more challenging to draw comparisons.

As can be seen in Table 5 speakers playing the role of the operator produce more acknowledgements in total than those playing the role of the caller: 36% as opposed to 26% of all utterances produced by each speaker. This can be attributed to the fact that the operator's task involves perceiving and understanding of names enquired by the caller, while the caller's task requires them to distinguish number sequences uttered by the operator. More specifically, this is due to the greater possibility for error in the understanding of names compared to numbers.

	Role					
Type	**Caller**		**Operator**		**Total**	
Ack	559	26%	726	36%	1285	31%
CR	94	4%	183	9%	277	7%
C	189	9%	64	3%	253	6%
Other	1306	61%	1044	52%	2350	56%
Total	**2148**	**100%**	**2017**	**100%**	**4165**	**100%**

Table 5: Results by speaker role and type of feedback

Our results show that the pattern of feedback mirrors the asymmetry of roles mentioned above, whereby the participant receiving the information is providing the majority of both acknowledgements and clarification requests (see Tables 6 and 7). Thus, 68% of utterances produced by the operators after a previous utterance containing some information about a business name are acknowledgements and 15% of them are clarification requests. Respectively, when it comes to the utterances produced by the callers following an utterance containing some information about a phone number, 73% of such utterances are acknowledgements and 12% of them are clarification requests.

This is consistent with the results reported in Colman & Healey (2011) and Rieser & Moore (2005) as pertaining to the distribution of clarification requests according to the role of the speaker in the context of

[4] The main findings have additionally been reported in Howes et al. (2019)

the task and suggests a similar pattern of distribution for acknowledgements.

	Role					
Type	**Caller**		**Operator**		**Total**	
Ack	50	11%	441	68%	491	44%
CR	3	1%	100	15%	103	9%
C	78	16%	1	0%	79	7%
Other	342	72%	105	16%	447	40%
Total	**473**	**100%**	**647**	**100%**	**1120**	**100%**

Table 6: Results by speaker role where the previous utterance is about a business name

	Role					
Type	**Caller**		**Operator**		**Total**	
Ack	364	73%	92	28%	456	55%
CR	60	12%	0	0%	60	7%
C	0	0%	30	9%	30	4%
Other	75	15%	210	63%	285	34%
Total	**499**	**100%**	**332**	**100%**	**831**	**100%**

Table 7: Results by speaker role where the previous utterance is about a business phone number

Our annotation and analysis results demonstrate that the speakers use different sub-types of both acknowledgements and clarification requests. As for the acknowledgements, these feedback utterances can combine several such sub-types in different order (Table 8). The most frequent in our corpus are those acknowledgements that contain at least one continuer (772 acknowledgements or 60%). The next most common way of acknowledging a preceding utterance in the dialogues is to repeat that whole utterance or a part of it verbatim (492 acknowledgements or 38%). The prevalence of these two acknowledgment types is favourable in terms of their implementation within dialogue systems as they constitute simple utterances that contain a continuer or a repetition. Interestingly, a combination of a verbatim repetition followed by a continuer is more frequent in the dialogues as opposed to a combination where a continuer is followed by a verbatim repetition. So, according to this, such an acknowledgement as “london okay” is more likely to be used by a dialogue participant than “okay london”.

Among the four clarification request sub-types distinguished by us (Table 9) confirmation requests constitute almost half of all CR utterances (134 CRs or 48%). These are used to check if some piece of information has been understood correctly by a participant. The majority of such CRs are very similar to verbatim repetitions discussed above, however, in this case they are produced with a rising intonation indicating uncertainty (Example 3, line 17).

The second most common way of clarifying a previous utterance in the dialogues (64 CRs or 23%) is through a repeat request (Example 3, line 11). This supports the claim made in Rieser & Moore (2005) which states that the participants initiating a clarification request tend to do so by asking their partner to confirm an existing hypothesis they have about an utterance rather than by prompting them to repeat the whole proposition.

As can be seen in Tables 10 and 11, clarification requests followed by clarifications are more likely to occur in the contexts where business names are discussed (8% for CRs and 11% for Cs) as opposed to the contexts where the speakers are discussing their phone numbers (5% for CRs and 5% for Cs). This indicates a greater chance for miscommunication and a higher incidence of repair events for the transmission of names rather than numbers. Within the domain of directory enquiries this mostly concerns the operator’s side of the

Type(s)	**Number**	**%**
Appreciate	5	0,4%
Confirm	21	1,6%
Confirm, Continuer	1	0,1%
Continuer	718	55,9%
Continuer, Appreciate	9	0,7%
Continuer, Appreciate, Continuer	1	0,1%
Continuer, Confirm	9	0,7%
Continuer, Paraphrase	2	0,2%
Continuer, Verbatim	3	0,2%
Paraphrase	25	1,9%
Paraphrase, Continuer	2	0,2%
Verbatim	456	35,5%
Verbatim, Appreciate	1	0,1%
Verbatim, Continuer	25	1,9%
Verbatim, Continuer, Appreciate	2	0,2%
Verbatim, Paraphrase	1	0,1%
Verbatim, Verbatim	4	0,3%
Total	**1285**	**100%**

Table 8: Types of acknowledgements

Type	**Number**	**%**
Confirmation request	134	48,4%
General request	28	10,1%
Repeat request	64	23,1%
Spelling request	51	18,4%
Total	**277**	**100%**

Table 9: Types of clarification requests

conversation, since they are the ones attempting to comprehend the enquired names. As it is essential for dialogue systems to be able to deal with such contexts, we now focus on the cases in our corpus where the feedback follows an utterance whose content is about a name.

As shown in Table 12, 45% of the turns that follow an utterance about a business name contain a complete spelling installment or a part of one, with similar proportions for acknowledgements (36%) and clarification requests (41%). Only 15% of turns contain a mention of a word or a word part belonging to the name in this position, with 12% being acknowledgements and 21% being clarification requests.

This suggests that the majority of such turns do not relate to the word level, but rather to the level of single letters or sequences thereof. From an implementation standpoint that means that the models of dialogue need to be able to produce and interpret increments of different sizes – potentially of a single letter, as people do when they are pinpointing sources of (potential) trouble within an unfamiliar name.

According to Tables 13 and 14, it is evident that the choice of a feedback strategy in the dialogues depends to a large extent on the strategy employed by the participant relaying the information in question in the preceding utterance. While it is typical of the feedback initiators to rely on generic strategies that do not involve verbatim or paraphrase repetitions of names in question (e.g. appreciative responses such as “great” or general clarification requests such as “pardon?”), it is also common for them to match the prior strategy

Type	Caller		Operator		Total	
Ack	249	66%	2	0%	251	30%
CR	41	11%		0%	41	5%
C		0%	38	8%	38	5%
Other	88	23%	414	91%	502	60%
Total	**378**	**100%**	**454**	**100%**	**832**	**100%**

Table 10: Types of utterances in the context of discussing business phone numbers

Type	Caller		Operator		Total	
Ack	9	1%	255	62%	264	24%
CR	4	1%	83	20%	87	8%
C	122	17%		0%	122	11%
Other	571	81%	76	18%	647	58%
Total	**706**	**100%**	**414**	**100%**	**1120**	**100%**

Table 11: Types of utterances in the context of discussing business names

	Ack		CR		Total	
Spelling installment	137	28%	31	30%	394	35%
Spelling installment part	41	8%	11	11%	107	10%
Word	21	4%	5	5%	47	4%
Word part	40	8%	16	16%	127	11%
Other	253	52%	42	41%	452	40%
Total	**491**	**100%**	**103**	**100%**	**1120**	**100%**

Table 12: Strategies for feedback following an utterance about a business name

during their turn. Thus, utterances that contain spelling installments are more likely to be acknowledged by spelling installments (40%), for example, while utterances that contain names pronounced in the form of a word are rarely sought to be clarified by a spelling installment (4%), and so on.

	Previous utterance content type								
	Sp. inst.		**Sp. inst. part**		**Word**		**Word part**		**Total**
Spelling installment	127	40%	9	20%		0%	1	1%	137
Spelling installment part	23	7%	18	39%		0%	4	6%	41
Word	3	1%	2	4%	10	20%	6	9%	21
Word part	3	1%		0%	15	30%	22	32%	40
(Continuer/Confirm/Appreciate)	171	54%	18	39%	25	50%	42	62%	253
Total	**319**	**100%**	**46**	**100%**	**50**	**100%**	**68**	**100%**	**491**

Table 13: Strategies for acknowledgements about a business name by previous utterance content type

	Previous utterance content type								
	Sp. inst.		**Sp. inst. part**		**Word**		**Word part**		**Total**
Spelling installment	24	52%	3	43%	1	4%	4	19%	31
Spelling installment part	8	17%	3	43%		0%		0%	11
Word		0%		0%	4	16%		0%	5
Word part	2	4%	1	14%	5	20%	10	48%	16
(General request/Repeat request)	17	37%		0%	16	64%	12	57%	42
Total	**46**	**100%**	**7**	**100%**	**25**	**100%**	**21**	**100%**	**103**

Table 14: Strategies for clarification requests about a business name by previous utterance content type

4.2 Qualitative results

Examples 4–7 show some of the feedback strategies in action. In Examples 4 and 6 the operators rely almost exclusively on continuer acknowledgements in their grounding of spelling installments. This, according to Clark & Schaefer (1989), represents a weaker evidence of understanding than, for example, verbatim repetitions (as in Example 4, lines 184–185, for example) and is more likely to result in misunderstandings. Examples 5 and 7 show the same business names as in the previous two excerpts split into two parts with the first part treated as an independent word ("hawk", "bistro") and the rest spelled out in installments of different size. It is noteworthy that different sub-parts of the spelled-out name in Example 7 are acknowledged in two different ways: there is a continuer at line 126 and a verbatim repetition at line 128.

Moreover, what is interesting about the exchange in Example 5 is the operator's tendency to respond with paraphrased displays of the name, where line 65 is acknowledging it while line 70 is pronounced with a rising intonation, making it a clarification request seeking to resolve the apparent problematic part of the utterance. This part, represented now as a spelling installment, is then repeated verbatim by the operator following a clarification from the caller at line 72.

(4) DEC24:169–193

169	Caller	it's a restaurant
170	Operator	yes
171	Caller	and it's called hawksmoor spi- spitalfields
172		so i'm gonna go ahead and- and spell it out
173	Operator	yes please
174	Caller	so H
175	Operator	mmhm

176	Caller	A
177	Operator	mmhm
178	Caller	W
179	Operator	yes
180	Caller	K
181	Operator	yes
182	Caller	S
183	Operator	mmhm
184	Caller	M
185	Operator	M
186	Caller	O
187	Operator	mmhm
188	Caller	O again
189	Operator	yes
190	Caller	R
191	Operator	yes
192	Caller	so try to - to see if you can find it with that
193	Operator	that’s all okay

(5) DEC05:60–75

60	Caller	and then it’s uh
61		restaurant
62		called er hawksmoor
63	Operator	hawk-
64	Caller	spitalfields
65	Operator	hawkswore spitalfields
66	Caller	hawksmoor
67		hawk you know like er like a bird
68	Operator	yeah
69	Caller	and then er
70	Operator	more M O R E ?
71	Caller	no it’s uh hawk S M O O R
72	Operator	M O O R
73		and spitalfield?
74	Caller	yes
75	Operator	one second

(6) DEC11:88–98

88	Operator	er can you spell bistrotheque for me?
89	Caller	abs-
90		sure er it’s
91		B I S
92	Operator	yes
93	Caller	T R O
94	Operator	mmhm
95	Caller	T H E
96	Operator	okay
97	Caller	Q U E

98	Operator	er yes i have it here for you

(7) DEC03:123–130

123	Caller	so bistro
124		T
125		H E
126	Operator	yeah
127	Caller	Q U E
128	Operator	Q U E
129	Caller	yeah
130	Operator	let me see if i find this

The data makes it possible to investigate how dialogue participants divide their presentations into what Clark & Schaefer (1987) refer to as installments, i.e. increments of spelled-out words or sub-parts of words, to aid understanding. Examples 4–7 feature various strategies in this regard ranging from one-letter utterances to installments of three letters. In Example 8 the caller splits the enquired name into two installments of three and four letters respectively, which the operator acknowledges with verbatim displays in each case.

(8) DEC16:54–61

54	Caller	the next place i'm looking for is called
55		er tayyabs which is spelled
56		T A Y
57	Operator	T A Y
58	Caller	Y A B S
59	Operator	Y A B S
60	Caller	it's a restaurant
61	Operator	okay

Thus, it is evident that the size of an installment is not fixed and that an installment might even be further subdivided in case of failure of understanding. This is demonstrated in Example 5 where the caller after an unsuccessful attempt at communicating the name "hawksmoor" splits it further into meaningful parts hoping that that would make the utterance less ambiguous ("hawk you know like er like a bird").

Examples 9 and 10 show two different ways the same name has been conveyed through spelling installments by two different pairs of participants. It took many more utterances for the pair in Example 10 to achieve the same goal, including the use of several verbatim acknowledgements. Additionally, in one of these acknowledgement cases the operator not just repeats the preceding installment but rather repeats the already seemingly grounded portion of the name up to that point in order to ensure a complete understanding (Example 10, line 110).

(9) DEC07:89–98

89	Caller	phoenicia mediterranean food[5]
90	Operator	can you repeat that for me?

[5] The exact pronunciations of the names featured in the dialogue examples as well as their mispronounced/misrendered variants can be found in the audio files in the published corpus (osf.io/2vjkh)

91		tenicia?
92	Caller	yeah
93		it's P H
94		O E N
95	Operator	mmhm
96		co- continue please
97	Caller	I C I A
98	Operator	I C I A

(10) DEC23:101–117

101	Caller	yeah it's phoenicia
102	Operator	clomissia?
103	Caller	mediterranean food
104		yes you spell it with a P
105	Operator	P
106	Caller	H
107		O
108	Operator	H O
109	Caller	E
110	Operator	yes P H O E
111	Caller	E N
112	Operator	N
113	Caller	A C
114	Operator	A C
115	Caller	A-
116		I A
117	Operator	I A

A common strategy for aiding the spelling process and thus avoiding misunderstandings is found in a number of dialogues in our corpus. This strategy usually involves providing the other participant with unambiguous words that have the same initials as the letters in a spelling installment. Different pairs come up with different categories of such words, with country/city names or people's first names being the most frequent choice. In Example 11 the caller is prompted to adopt this strategy by a clarification request from the operator (line 19), who then in their acknowledgements of such spelling installments either drops the letter and only repeats the non-ambiguous word (lines 23, 30, 32) or repeats the whole utterance (line 35).

The disambiguation strategy can be initiated by either participant and there is some indication in our data that after repeated interactions participants may start to mirror each other's spelling style, where one of them will adopt the other one's strategy or perform it in co-construction with each other as in Example 11, lines 138–141. Furthermore, there are cases where the letter in question is not pronounced leaving the unambiguous word to stand in for the whole (Example 12).

(11) DEC28:17–141

17	Caller	okay so it starts with a
18		L
19	Operator	L?
20	Caller	as in london
21	Operator	yes

22	Caller	A as in america
23	Operator	america
24	Caller	er U
25		as in er
26		er under
27		<laugh>
28	Operator	under yes
29	Caller	er D as in denmark
30	Operator	denmark
31	Caller	E as in england
32	Operator	england
33	Caller	and R
34		for russia
35	Operator	R for russia
:	:	:
138	Caller	and K for er
139		<laugh>
140	Operator	as in king?
141	Caller	k- king <laugh> yeah

(12) DEC26:61–69

61	Caller	it's it's a restaurant by name tayyabs
62	Operator	okay can you spell that for me please?
63	Caller	should i
64		yes it's a thailand
65	Operator	yes
66	Caller	america
67	Operator	yes
68	Caller	yugoslavia
69	Operator	yes
:	:	:

Interestingly, when it comes to non-conventional spellings of words there is some indication in our data that participants are generally good at predicting potentially problematic elements. In some cases such elements are pinpointed and specified before they lead to miscommunication. Examples 13 and 14 both provide illustrations of this particular strategy of successfully averting potential failures.

(13) DEC20:4–9

4	Caller	the first one being first one being one called cittie of yorke which is C I T T I E of
5		yorke spelled with an E at the end
6	Operator	cittie of yorke with two Ts?
7	Caller	cittie of yorke where cittie isn't
8		C I T Y it's C I T T I E
9	Operator	yeah

(14) DEC10:59–61

59	Caller	it's called lyle's
60		with a Y
61	Operator	lyle's

In case any misunderstandings do arise, the data suggests that they are largely resolved quickly and locally, however, there are some interesting cases where they persist. In Example 15, where a specific problematic letter in the name takes 57 utterances to resolve, the caller starts by spelling out the enquired name and then eventually has to change their strategy to the one that utilises initial letters of first names. After several unsuccessful attempts, including going through several different first names, the problematic letter is resolved by using a common unambiguous word instead of a first name (line 136). This example shows how a widely used spelling out strategy can in itself become a source of miscommunication, especially in noisier environments.

(15) DEC22:82–139

82	Caller	with a - filip with an F
83	Operator	filip
84		yeah
:	:	:
107	Caller	er
108	Operator	pilip
109	Caller	fanny
110	Operator	mmhm
111	Caller	fanny
:	:	:
113	Operator	P
114		P as in panda
115		right?
116	Caller	sorry i didn't hear you
117	Operator	P
118		the next one is a P
119		as in panda
120	Caller	P?
121		or okay
122		then
123	Caller	no
124		it's er
:	:	:
133	Caller	uh fanny
134	Operator	<unclear> I don't know that name funny?
135	Caller	yeah or like filip but with an F
136		or if you say fruits
137	Operator	with an F?
138		okay
139	Caller	F yeah

There are few (3 out of 84[6]) cases of complete failure to communicate and ground the enquired information in our corpus. In Example 16 the miscommunication stems from the similarity in sound of a “B” and a “V” (especially for the caller who is a native speaker of Spanish) as well as the noisy environment interference. This case does not get resolved and has to be abandoned after multiple back-and-forth attempts at spelling and requests for clarification. It bears noting that unlike the situation in Example 15, the participants in this particular conversation have not been able to identify the source of the problem, which could have been avoided or resolved had they opted for the first name/place name strategy. With this in mind, it is, therefore, crucial for dialogue systems to be able to generate or prompt the user for such strategies, especially in cases where the user does not initiate them on their own.

(16) DEC14:4–112

4	Caller	er one is a pub
5		it’s called the star tavern
6	Operator	can you repeat please?
7	Caller	the star
8		tavern
:	:	:
29	Operator	i’m not sure if i heard the name of the place correctly
30		can you repeat?
31	Caller	yeah the the name of the place the
32	Operator	yes
33	Caller	the tavern it’s the star
34		star like a star in the sky you know <laugh>
35	Operator	yes
36	Caller	the night
37	Operator	mmhm
38	Caller	er tavern
39	Operator	can you spell it er please ta-?
:	:	:
58	Caller	and it’s tavern it’s T A
59	Operator	and then
60	Caller	V E er <R> un <N>
61		N
62		sorry
:	:	:
72	Operator	T A B E R N
73		is that correct?
74	Caller	yeah
:	:	:
96	Operator	star tabern right?
97	Caller	yeah
:	:	:
112	Operator	website still says we’re sorry we co- couldn’t find any results

[6] In total, there are 84 instances of name grounding in the corpus

4.3 Finite-state model of grounding names

As the next step in our analysis we present a tentative formalisation of the process of grounding of names following Traum's descriptive finite-state model of grounding in dialogue (Traum, 1994). The result is a finite-state network that makes it possible to track the state of the dialogue with regard to the grounding process between the point where the caller first mentions an enquired name and the point where the operator signals that name as having been grounded (see Figure 2 below). The model was created by analysing the relevant portions of the annotated data and investigating recurrent patterns of feedback. It is important to note that the suggested model is part of a preliminary effort to formalise name grounding based on our observations about the available data and does not constitute a definitive representation of the process.

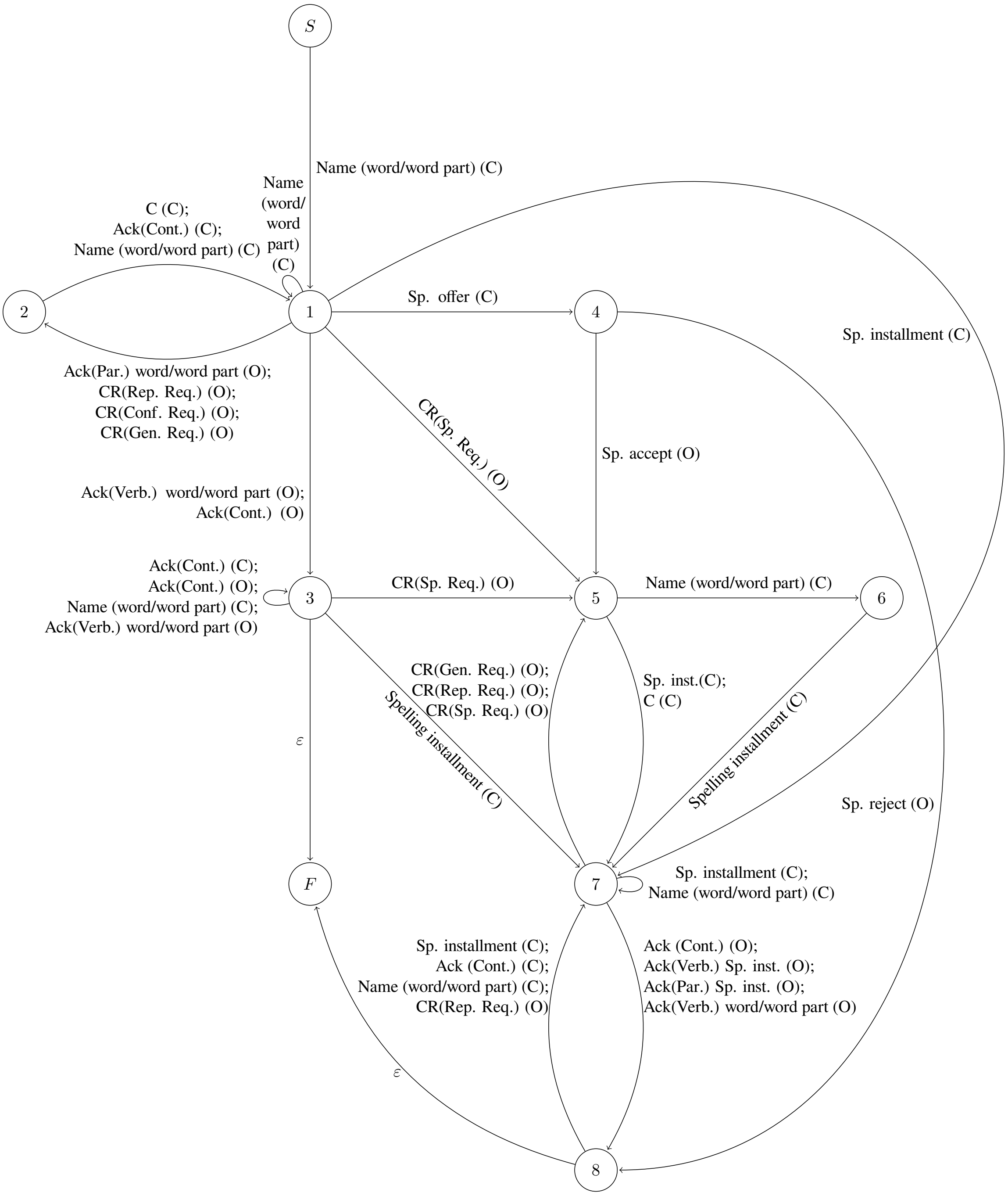
The network has 10 possible states with states S and F, similarly to Traum's model, representing respectively a state where a proposition (specifically a name in our case) has not been initiated yet and a state where it has been grounded. The transition from state S to state 1 corresponds to the first mention of the name or part of the name by the caller, which can sometimes be followed by a self-repetition or a continuation of the enquiry, hence there is a possibility for recursion in this state. Transitions from state 1 to state 2 and back represent a potential exchange between the participants where the operator either fails at correctly repeating the name in question resulting in a paraphrase acknowledgement (Example 17, line 65) or initiates a clarification request (a confirmation request, for example, as in Example 18, lines 63 and 65). The caller's response in this case is either a clarification (Example 18, lines 64 and 66), a continuer acknowledgement or a restatement of the material (Example 17, line 66).

One of the examples of a minimum number of transitions between states to reach the grounded state is Example 19, where at state 1 what was needed to reach state F was a verbatim acknowledgement coming from the operator followed by a continuer from the caller and a final continuer from the operator. However, there is a possibility the dialogue will have to go into a spelling out phase at this particular point. This can be done in three different ways: either through a caller's initiation of a spelling installment which would take the dialogue directly into state 7 (Example 17, line 66), a spelling request by the operator (Example 20, line 16) or a spelling offer - spelling accept transition (Example 21, lines 6–7) both of which would take the dialogue into state 5.

State 5 in the network represents a point where a spelling installment can either be initiated directly and transition to state 7 (Example 21, line 8) or it can be initiated after a name restatement by the caller in which case the dialogue moves into state 7 through state 6. What follows is essentially a spelling out loop (7–8–7) where the most frequent transition is through a verbatim/continuer acknowledgement by the operator and a new spelling installment by the caller (Example 17, lines 67–78; Example 20, lines 19–41; Example 21, lines 9–13).

Furthermore, the state 8 to state 7 transition includes some other options such as a continuer acknowledgment by the caller (Example 17, line 78; Example 21, line 14), a name repetition or initiation (Example 17, line 86), or a repetition request by the operator. The response to that can contain a paraphrase of a spelling installment or a verbatim acknowledgement of a name by the operator (Example 17, lines 83 and 85). Additionally, the 7–7 recursion allows for spelling installment or name repetitions (Example 17, lines 79–80), as well as spelling initiations in cases where the previous transition included a clarification from the caller as a response to a spelling request from the operator at states 1–5–7 or 3–5–7 (Example 20, lines 16–18).

Finally, in addition to the epsilon transition from state 3 and depending on the context as well as the current state, state F can be reached through an epsilon transition from state 8 (Example 17, line 87; Example 20, line 41; Example 21, line 15). Interestingly, our data shows that the most frequent final move at the end of the name grounding process is the one that includes a continuer acknowledgement by the operator (specifically “okay” as in Examples 18, 19 and 21). This is consistent with the notion of a “horizontal vs. vertical” transition between parts of the dialogue discussed in Bangerter & Clark (2003), wherein the continuer would be a marker of vertically initiating a closing of the name grounding sequence and moving into another stage of the conversation.



C = Clarification
Ack(Cont.) = Acknowledgement (Continuer)
Ack(Par.) = Acknowledgement (Paraphrase)
Ack(Verb.) = Acknowledgement (Verbatim)
CR(Rep. Req) = Clarification request (Repeat request)
CR(Conf. Req) = Clarification request (Confirmation request)
CR(Gen. Req) = Clarification request (General request)
CR(Sp. Req) = Clarification request (Spelling request)
Sp. inst. = Spelling installment
Sp. offer, Sp. accept, Sp. reject = Spelling offer/accept/reject

Figure 2: Finite-state network for grounding names

(17) DEC01:64–87

64	Caller	uh er the place is er todich	Name (word/word part) (C)	S-1
65	Operator	todit uh i	Ack(Par.) word/word part (O)	1-2
66	Caller	uh todich is like er T for er er thailand	Name (word/word part) (C) + Sp. installment (C)	2-1; 1-7
67	Operator	T	Ack(Verb.) Sp. inst. (O)	7-8
68	Caller	uh O for	Sp. installment (C)	8-7
69	Operator	O	Ack(Verb.) Sp. inst. (O)	7-8
70	Caller	D for denmark	Sp. installment (C)	8-7
71	Operator	D	Ack(Verb.) Sp. inst. (O)	7-8
72	Caller	I for india	Sp. installment (C)	8-7
73	Operator	I	Ack(Verb.) Sp. inst. (O)	7-8
74	Caller	C for canada	Sp. installment (C)	8-7
75	Operator	C for canada	Ack(Verb.) Sp. inst. (O)	7-8
76	Caller	H for hawaii	Sp. installment (C)	8-7
77	Operator	H for hawaii	Ack(Verb.) Sp. inst. (O)	7-8
78	Caller	yeah	Ack (Cont.) (C)	8-7
79		and er it's er flo- floral	Name (word/word part) (C)	7-7
80		the second word of that company is floral	Name (word/word part) (C)	7-7
81	Operator	floral	Ack(Verb.) word/word part (O)	7-8
82	Caller	yeah	Ack (Cont.) (C)	8-7
83	Operator	todich floral	Ack(Verb.) word/word part (O)	7-8
84	Caller	design	Name (word/word part) (C)	8-7
85	Operator	design	Ack(Verb.) word/word part (O)	7-8
86	Caller	yeah so todich floral design	Name (word/word part) (C)	8-7
87	Operator	yes i got it here	Ack(Cont.)(O)	7-8-F

(18) DEC07:62–68

62	Caller	and er the next number i'm looking for is the peasant	Name (word/word part) (C)	S-1
63	Operator	the peasant?	CR(Conf. Req.) (O)	1-2
64	Caller	yes	C (C)	2-1
65	Operator	like in uh farmer?	CR(Conf. Req.) (O)	1-2
66	Caller	exactly	C (C)	2-1
67	Operator	the peasant	Ack(Verb.) word/word part (O)	1-3
68		okay	Ack(Cont.) (O)	3-3-F

(19) DEC01:106–109

106	Caller	yeah er the name of the place it's sweet things	Name (word/word part) (C)	S-1
107	Operator	sweet things	Ack(Verb.) word/word part (O)	1-3
108	Caller	yeah exactly	Ack(Cont.) (C)	3-3
109	Operator	okay	Ack(Cont.) (O)	3-3-F

(20) DEC25:15–42

15	Caller	silver cross	Name (word/word part) (C)	S-1

16	Operator	can you spell that one for me please?	CR(Sp. Req.) (O)	1-5
17	Caller	yes	C (C)	5-7
18	Caller	S	Sp. installment (C)	7-7
19	Operator	S	Ack(Verb.) Sp. inst. (O)	7-8
20	Caller	I	Sp. installment (C)	8-7
21	Operator	I	Ack(Verb.) Sp. inst. (O)	7-8
22	Caller	L	Sp. installment (C)	8-7
:	:	:	:	:
40	Caller	S	Sp. installment (C)	8-7
41	Operator	S	Ack(Verb.) Sp. inst. (O)	7-8-F

(21) DEC21:5–15

5	Caller	that i need to find the fi-first is er chesneys	Name (word/word part) (C)	S-1
6		uh do you want me to spell it?	Sp. offer (C)	1-4
7	Operator	yes please	Sp. accept (O)	4-5
8	Caller	uh C H	Sp. installment (C)	5-7
9	Operator	yes	Ack(Cont.) (O)	7-8
10	Caller	E S	Sp. installment (C)	8-7
11	Operator	yeah	Ack(Cont.) (O)	7-8
12	Caller	N E Y S	Sp. installment (C)	8-7
13	Operator	N E Y S	Ack(Verb.) Sp. inst. (O)	7-8
14	Caller	yes	Ack(Cont.) (C)	8-7
15	Operator	okay	Ack(Cont.) (O)	7-8-F

5 Conclusions and future work

In this thesis we have presented a new corpus of dialogues in the domain of directory enquiries and described the process of its collection, transcription and annotation. The published corpus consists of 28 dialogues including 84 instances of name grounding (osf.io/2vjkh; Bondarenko et al., 2019). The focus of the annotation process was on three main types of utterances which we tagged as acknowledgements (Ack), clarification requests (CR) and clarifications (C), with the first two representing feedback utterances and being further annotated into sub-types. Specifically, for acknowledgements the sub-types were continuer, verbatim, paraphrase, confirm and appreciate, while for clarification requests they included general requests, repeat requests, confirmation requests and spelling requests. A number of additional tags were assigned to each utterance in order to assist with future exploration of the data (see Section 3.2).

Additionally, we conducted a preliminary investigation of feedback strategies found in the corpus and discussed the related implications for dialogue systems. In Section 4.1 we elaborated on the quantitative results of the corpus study which demonstrated a higher degree of feedback use than reported in most of the previous work (37%). Furthermore, it was established that the dialogue participants playing the role of the operator produced more acknowledgements than those playing the role of the caller (36% vs. 26%) which might be due to the greater possibility of miscommunication in the transmission of names rather than number sequences. As for the different sub-types of feedback utterances, continuers appeared to be the most frequent one (60%) with verbatim repetitions being the second most common type (38%). Confirmation requests were found to constitute almost half of all clarification request utterances (48%) while repeat requests were found to constitute 23% of all CRs. Moreover, according to our analysis, feedback strategies tend to mirror the choice of strategy used to relay the information in the preceding utterance (spelling installments are more likely to be acknowledged by spelling installments (40%) and so on).

Section 4.2 was dedicated to the discussion of qualitative results of the corpus study where we provided a number of dialogue examples to illustrate the use of specific feedback strategies by the participants. Namely, these examples demonstrated the ways the participants divide their presentations into installments of a non-fixed size and how different sub-parts of spelling installments can be acknowledged in different ways within the same exchange. Beyond that, the dialogue excerpts showcased a common disambiguation strategy used by a large number of speakers, as well as how the speakers normally resolve misunderstandings, as we argued that it is essential for dialogue systems dealing with communication of accurate information between dialogue partners to be able to generate or prompt the user for such strategies.

Finally, in Section 4.3 we proposed a tentative finite-state model of grounding of names following Traum's descriptive model of grounding (Traum, 1994). The model is based on the observations we have made regarding recurrent patterns of feedback in the collected data. It has 10 possible states and covers the cases where it takes a minimum number of transitions to reach the grounded state, as well as those ones where the participants have to initiate a spelling out phase, which can be done either through a direct spelling installment initiation, spelling offer - spelling accept transition, or a spelling request by the operator (Figure 2).

As one of the main contributions of this thesis was the collection and publication of a directory enquiries corpus, and since the size of the corpus is somewhat limited due to the scope of the project, it seems auspicious to extend the corpus in the future placing emphasis on inter-rater reliability measurement for the parts of the annotation where it has not yet been established. Additionally, as previously mentioned (see Section 3.2), it might be of significance to reconsider the feedback sub-type classification suggested by us in a way that would not conflate form and function (Howes et al., 2019), as well as to possibly distinguish additional types of feedback utterances and their combinations.

Apart from that, since the primary focus of our corpus study was on grounding instances which concerned discussions about business names as opposed to instances where phone numbers were discussed it could be interesting to investigate the latter contexts in more detail and compare findings, especially in regards to the choice of feedback strategies and division of such material into installments. Likewise, it would be logical to extend our grounding network to accommodate the grounding of number sequences and possibly make it a probabilistically weighted one. The proposed model would also need to be evaluated against more data to establish the extent of its coverage as well as considered in the context of existing implementations of models of grounding within incremental dialogue systems (Skantze & Schlangen, 2009; Buß et al., 2010; Visser et al., 2014; Wang et al., 2011; Khouzaimi et al., 2014). Specifically, the aforementioned work on formal models of grounding (including also Traum, 1994 and Larsson, 2002) has often assumed words to be the minimal grounded unit. However, in a complete dialogue model that deals with accuracy-sensitive content grounding of words needs to be combined with grounding of their sub-parts including sometimes single letters, as demonstrated by the grounding behaviours discussed in this thesis.

References

Allwood, J. (1999). The Swedish spoken language corpus at Göteborg University. In *Fonetik '99, Gothenburg Papers in Theoretical Linguistics 81*.

Anderson, A. H., Bader, M., Bard, E. G., Boyle, E., Doherty, G., Garrod, S., Isard, S., Kowtko, J., McAllister, J., Miller, J., Sotillo, C., Thompson, H. S., & Weinert, R. (1991). The HCRC map task corpus. *Language and Speech*, 34(4), 351–366.

Bangerter, A. & Clark, H. H. (2003). Navigating joint projects with dialogue. *Cognitive Science*, 27, 195–225.

Bell, L. & Gustafson, J. (2000). Positive and negative user feedback in a spoken dialogue corpus. In *INTERSPEECH 2000*.

Bondarenko, A., Howes, C., & Larsson, S. (2019). Directory Enquiries Corpus. Available at https://osf.io/2vjkh/.

Boyle, E. A., Anderson, A. H., & Newlands, A. (1994). The effects of visibility on dialogue and performance in a cooperative problem solving task. *Language and Speech*, 37(1), 1–20.

Brugman, H. & Russel, A. (2004). Annotating multi-media/multi-modal resources with ELAN. In *LREC*.

Burnard, L. (2000). Reference guide for the British National Corpus (World Edition).

Buß, O., Baumann, T., & Schlangen, D. (2010). Collaborating on utterances with a spoken dialogue system using an ISU-based approach to incremental dialogue management. In *Proceedings of the 11th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, SIGDIAL '10 (pp. 233–236). Stroudsburg, PA, USA: Association for Computational Linguistics.

Cahn, J. E. & Brennan, S. E. (1999). A psychological model of grounding and repair in dialog. In *Proc. Fall 1999 AAAI Symposium on Psychological Models of Communication in Collaborative Systems*.

Cerrato, L. (2002). A comparison between feedback strategies in human-to-human and human-machine communication. In *INTERSPEECH 2002*.

Chang, H. (2007). Comparing machine and human performance for caller's directory assistance requests. *International Journal of Speech Technology*, 10(2-3), 75–87.

Clark, H. H. & Schaefer, E. F. (1987). Collaborating on contributions to conversations. *Language and Cognitive Processes*, 2(1), 19–41.

Clark, H. H. & Schaefer, E. F. (1989). Contributing to discourse. *Cognitive Science*, 13(2), 259–294.

Colman, M. & Healey, P. G. T. (2011). The distribution of repair in dialogue. In *CogSci*.

Duncan, S. (1972). Some signals and rules for taking speaking turns in conversations. *Journal of Personality and Social Psychology*, 23(2), 283–292.

Duncan, S. (1974). On the structure of speaker-auditor interaction during speaking turns. *Language in Society*, 3(2), 161–180.

Fernández, R. (2006). *Non-Sentential Utterances in Dialogue: Classification, Resolution and Use*. PhD thesis, King's College London, London, UK.

Fujimoto, D. T. (2007). Listener responses in interaction: A case for abandoning the term, backchannel. *Journal of Osaka Jogakuin 2year College*, 37.

Hjalmarsson, A. & Edlund, J. (2008). Human-likeness in utterance generation: Effects of variability. In *Perception in Multimodal Dialogue Systems: 4th IEEE Tutorial and Research Workshop on Perception and Interactive Technologies for Speech-Based Systems, PIT 2008, Kloster Irsee, Germany, June 16-18, 2008. Proceedings (pp.252-255)*, volume 5078 (pp. 252–255).

Howes, C., Bondarenko, A., & Larsson, S. (2019). Good call! Grounding in a Directory Enquiries Corpus. In *Proceedings of the 23rd Workshop on the Semantics and Pragmatics of Dialogue - Full Papers* London, United Kingdom: SEMDIAL.

Khouzaimi, H., Laroche, R., & Lefèvre, F. (2014). An easy method to make dialogue systems incremental. In *SIGDIAL Conference*.

Koit, M. (2012). Towards automatic recognition of the structure of Estonian directory inquiries. *Frontiers in Artificial Intelligence and Applications*, 247, 120–128.

Larsson, S. (2002). *Issue-based Dialogue Management*. PhD thesis, Goteborg University.

Passonneau, R. & Sachar, E. (2014). Loqui human-human dialogue corpus (transcriptions and annotations).

Purver, M. (2004). *The Theory and Use of Clarification Requests in Dialogue*. PhD thesis, King's College London.

Purver, M., Howes, C., Healey, P. G. T., & Gregoromichelaki, E. (2009). Split utterances in dialogue: A corpus study. In *Proceedings of the SIGDIAL 2009 Conference: The 10th Annual Meeting of the Special Interest Group on Discourse and Dialogue* (pp. 262–271).: Association for Computational Linguistics.

Rieser, V. & Moore, J. (2005). Implications for generating clarification requests in task-oriented dialogues. In *43rd Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference*.

Rühlemann, C. (2007). *Conversation in Context: A Corpus-Driven Approach*. Continuum.

Schegloff, E., Jefferson, G., & Sacks, H. (1977). The preference for self-correction in the organization of repair in conversation. *Language*, 53, 361–382.

Schegloff, E. & Sacks, H. (1973). Opening up closings. *Semiotica*, 8, 289–327.

Skantze, G., House, D., & Edlund, J. (2006). User responses to prosodic variation in fragmentary grounding utterances in dialog. In *INTERSPEECH 2006*.

Skantze, G. & Schlangen, D. (2009). Incremental dialogue processing in a micro-domain. In *EACL 2009, 12th Conference of the European Chapter of the Association for Computational Linguistics, Proceedings of the Conference* (pp. 745–753).

Traum, D. R. (1994). *A Computational Theory of Grounding in Natural Language Conversation*. PhD thesis, University of Rochester, Rochester, NY, USA. UMI Order No. GAX95-23171.

van Heerden, C. J., Davel, M. H., & Barnard, E. (2014). Performance analysis of a multilingual directory enquiries application. In *Proceedings of the Annual Symp. Pattern Recognition Association of South Africa (PRASA)* (pp. 258–263).

Visser, T., Traum, D., DeVault, D., & Akker, R. (2014). A model for incremental grounding in spoken dialogue systems. *Journal on Multimodal User Interfaces*, 8.

Wang, Z., Lee, J., & Marsella, S. (2011). Towards more comprehensive listening behavior: Beyond the bobble head. In *Proceedings of the 10th International Conference on Intelligent Virtual Agents*, IVA'11 (pp. 216–227). Berlin, Heidelberg: Springer-Verlag.

Ward, K. & Heeman, P. A. (2000). Acknowledgments in human-computer interaction. In *Proceedings of the 1st North American Chapter of the Association for Computational Linguistics Conference*, NAACL 2000 (pp. 280–287). Stroudsburg, PA, USA: Association for Computational Linguistics.

Appendices

Appendix 1

Instructions for the Caller

In this setup you are playing the role of a telephone directory service Caller. You will be given a list of businesses located in London and your task is to find out their phone numbers. For this you are asked to call up the Operator at the number you will also be provided with. You will need to make 2 calls from 2 different locations. These conversations will be recorded. Please do not reveal any personal information about yourself or the other participant during the calls.

1. Go to the first location
2. Put on the recorder headset (assistant starts the recording)
3. Call the Operator
4. When the Operator answers the call, the person assisting you will do three sync claps, after which the Operator will do three more sync claps. After that, you can proceed with the conversation
5. Ask the Operator to give you the phone numbers for the places on your list and fill in the table with those numbers
6. Finish the call (assistant stops the recording)
7. Move to the second location and repeat steps 2-6

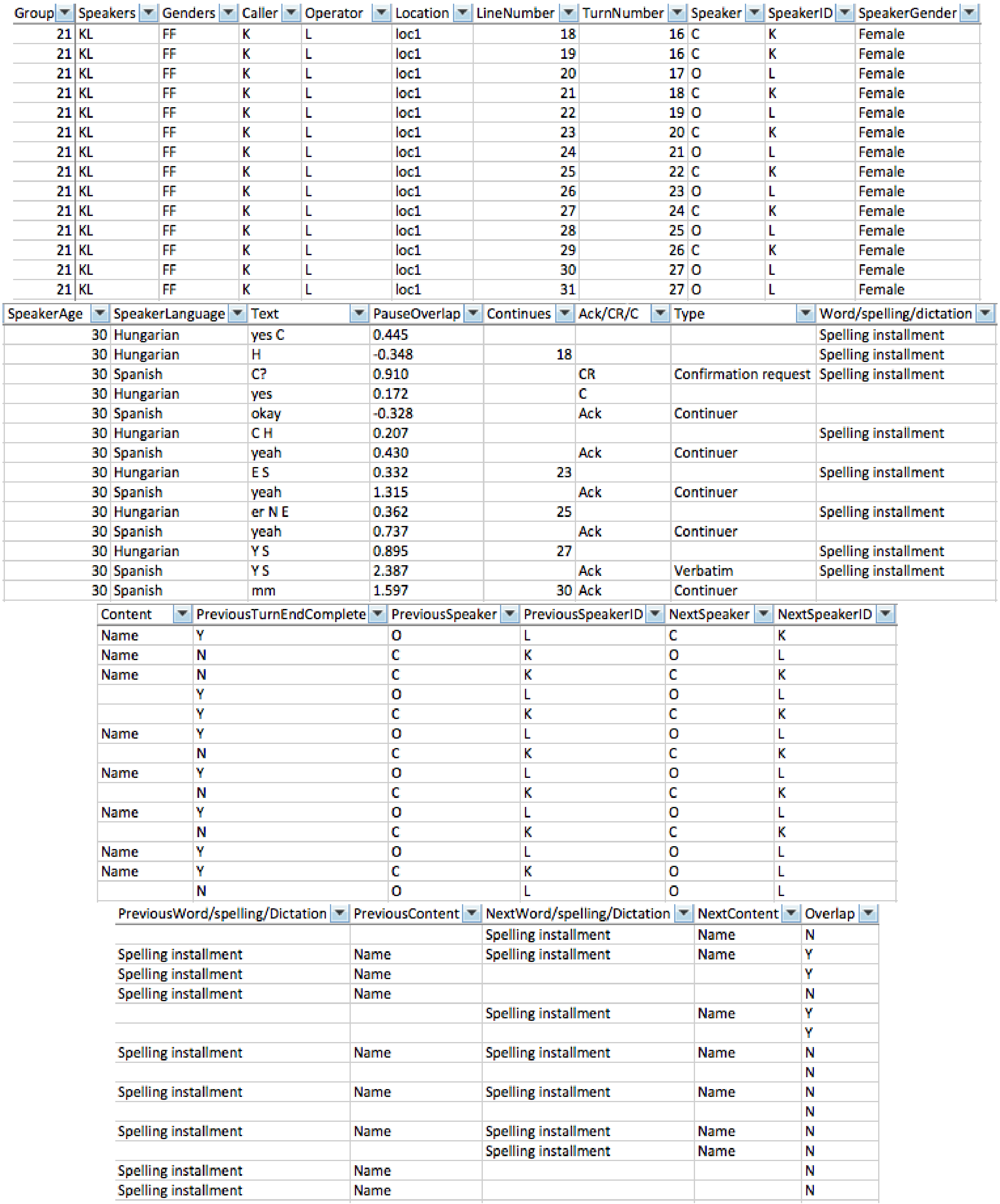
Appendix 2

Instructions for the Operator

In this setup you are playing the role of a telephone directory service Operator. You will be provided with a laptop and a phone. You will get 2 phone calls on that phone and you are asked to provide the Caller with the phone numbers of the places located in London they are interested in. You will do this with the help of the online Phone Book service (thephonebook.bt.com). These conversations will be recorded. Please do not reveal any personal information about yourself or the other participant during the calls.

1. When you get a call from the other participant, press the recording button on the studio computer
2. Take a seat in front of the laptop and receive the call
3. Wait for three sync claps on the other end of the call
4. Do three sync claps in the studio
5. Making sure that the search on thephonebook.bt.com is set to "Name of Business" and "in London", provide the Caller with the phone numbers of different businesses by looking them up in the Phone Book (everything they are asking about can be found in the Phone Book)
6. Once the Caller is satisfied, end the call
7. Stop the recording on the studio computer
8. Repeat steps (1-7) one more time once the other participant has reached the second location

Appendix 3



Group	Speakers	Genders	Caller	Operator	Location	LineNumber	TurnNumber	Speaker	SpeakerID	SpeakerGender
21	KL	FF	K	L	loc1	18	16	C	K	Female
21	KL	FF	K	L	loc1	19	16	C	K	Female
21	KL	FF	K	L	loc1	20	17	O	L	Female
21	KL	FF	K	L	loc1	21	18	C	K	Female
21	KL	FF	K	L	loc1	22	19	O	L	Female
21	KL	FF	K	L	loc1	23	20	C	K	Female
21	KL	FF	K	L	loc1	24	21	O	L	Female
21	KL	FF	K	L	loc1	25	22	C	K	Female
21	KL	FF	K	L	loc1	26	23	O	L	Female
21	KL	FF	K	L	loc1	27	24	C	K	Female
21	KL	FF	K	L	loc1	28	25	O	L	Female
21	KL	FF	K	L	loc1	29	26	C	K	Female
21	KL	FF	K	L	loc1	30	27	O	L	Female
21	KL	FF	K	L	loc1	31	27	O	L	Female

SpeakerAge	SpeakerLanguage	Text	PauseOverlap	Continues	Ack/CR/C	Type	Word/spelling/dictation
30	Hungarian	yes C	0.445				Spelling installment
30	Hungarian	H	-0.348	18			Spelling installment
30	Spanish	C?	0.910		CR	Confirmation request	Spelling installment
30	Hungarian	yes	0.172		C		
30	Spanish	okay	-0.328		Ack	Continuer	
30	Hungarian	C H	0.207				Spelling installment
30	Spanish	yeah	0.430		Ack	Continuer	
30	Hungarian	E S	0.332	23			Spelling installment
30	Spanish	yeah	1.315		Ack	Continuer	
30	Hungarian	er N E	0.362	25			Spelling installment
30	Spanish	yeah	0.737		Ack	Continuer	
30	Hungarian	Y S	0.895	27			Spelling installment
30	Spanish	Y S	2.387		Ack	Verbatim	Spelling installment
30	Spanish	mm	1.597	30	Ack	Continuer	

Content	PreviousTurnEndComplete	PreviousSpeaker	PreviousSpeakerID	NextSpeaker	NextSpeakerID
Name	Y	O	L	C	K
Name	N	C	K	O	L
Name	N	C	K	C	K
	Y	O	L	O	L
	Y	C	K	C	K
Name	Y	O	L	O	L
	N	C	K	C	K
Name	Y	O	L	O	L
	N	C	K	C	K
Name	Y	O	L	O	L
	N	C	K	C	K
Name	Y	O	L	O	L
Name	Y	C	K	O	L
	N	O	L	O	L

PreviousWord/spelling/Dictation	PreviousContent	NextWord/spelling/Dictation	NextContent	Overlap
		Spelling installment	Name	N
Spelling installment	Name	Spelling installment	Name	Y
Spelling installment	Name			Y
Spelling installment	Name			N
		Spelling installment	Name	Y
				Y
Spelling installment	Name	Spelling installment	Name	N
				N
Spelling installment	Name	Spelling installment	Name	N
				N
Spelling installment	Name	Spelling installment	Name	N
		Spelling installment	Name	N
Spelling installment	Name			N
Spelling installment	Name			N

Figure 3: Annotation example for DEC21:18–31