Could you spell that again please?
Towards a Formal Model of Grounding in Directory Enquiries

Staffan Larsson **Christine Howes** **Anastasia Bondarenko**
Centre for Linguistic Theory and Studies in Probability (CLASP)
Department of Philosophy, Linguistics and Theory of Science
University of Gothenburg, Sweden
staffan.larsson@gu.se

Abstract

Based on a corpus of directory enquiries dialogue and a preliminary analysis of dialogue act sequences involved in spelling out names in these dialogues, we provide the basic components of an incremental computational account of grounding of names. The account is inspired by previous work on a computational model for incremental processing that has previously been applied to number sequences. Here, we tackle the issue of the spelling out of names, where more complex behaviours are involved, mixing fully read out names, partially spelled names, clarification requests, and more. This model has a potential application in designing conversational agents that can handle previously unencountered names with possibly idiosyncratic spellings.

1 Introduction

Effective communication requires collaboration between all participants, with dialogue co-constructed by speakers and hearers. Even in contexts such as lectures or storytelling, which are largely monological (Rühlemann, 2007), listeners provide frequent feedback. This feedback demonstrates whether or not they have *grounded* the conversation thus far (Clark, 1996), i.e. whether something said can be taken to be understood. Positive feedback, indicating sufficient understanding comes in the form of relevant next turns, or backchannels (e.g. 'yes', 'yeah', Example 1; lines 6 and 8[1] or 'mm').[2] Other responses, such as clarification requests (e.g. Example 1; lines 10 and 17) indicate processing difficulties or lack of coordination and signal a need for repair (Purver, 2004; Bavelas et al., 2012).

These communicative grounding strategies (Clark and Brennan, 1991; Traum, 1994) enable dialogue participants to manage the characteristic divergence and convergence that is key to moving dialogue forward (Clark and Schaefer, 1987, 1989), and are therefore crucial for dialogue agents. Importantly, feedback is known to occur subsententially (Howes and Eshghi, 2017), but most dialogue models do not operate in an incremental fashion that would allow them to produce or interpret feedback in a timely fashion.

(1) 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
20	Operator	and pin?
21	Caller	yes
22	Caller	yes
23	Operator	tenpin
24	Operator	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

Here, we focus on feedback in an extremely restricted domain – that of telephone directory

[1] Examples are all taken from our Directory Enquiries Corpus (DEC), described below.

[2] In face-to-face dialogue this includes non-linguistic cues (e.g. nods), but as our corpus is telephone conversations, we do not consider these here.

enquiries (see also Clark and Schaefer, 1987; Bangerter et al., 2004), which can be seen as a good test case for dialogue systems. Directory enquiries is a real world application for dialogue systems (e.g. Chang, 2007) that has particular features that can be problematic for a speech recogniser, such as understanding names which are not present in an existing lexicon over a noisy channel. As we argue below, this is a particularly good domain for studying feedback, as feedback is more frequent and necessary than in less restricted domains. The reasons for this are two-fold. Firstly, in task-oriented dialogue, where information transfer is crucial for success, and avoiding miscommunication is vital, feedback is more common than in less goal-directed conversations (Colman and Healey, 2011). Secondly, verbal feedback is more frequent in dialogues where participants cannot see each other, and therefore do not have the ability to employ non-verbal feedback (Boyle et al., 1994), such as telephone conversations. Additionally, in contrast to corpora which have similar features (such as SRI's Amex Travel Agent Data, Kowtko and Price, 1989), relevant parts of the dialogue (such as names, see below) do not require anonymisation.

2 Previous work

In Howes et al. (2019), a freely available corpus of human-human telephone directory enquiries dialogues was presented, and the strategies for feedback that human participants use, especially in cases where misunderstandings arise, were explored. It was suggested that dialogue models need to be able to perform incremental grounding, particularly in the context of spelling out Names and dictating number sequences, with a number of increasingly specific strategies available for both acknowledgements and clarifications.[3]

Work on formal modelling of grounding (e.g. Traum, 1994; Larsson, 2002; Visser et al., 2014) has often assumed that the minimal units being grounded are words. In a complete model, this needs to be complemented by the grounding of subparts of words, including single letters. Work in this direction includes Skantze and Schlangen (2009), where dictation of number sequences is used as a test case "micro-domain" for an implemented model of incremental grounding. Continuing this work Buß et al. (2010) further extend the system's capabilities to accommodate grounding in semantically more complex domains where grounding on the understanding level 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).

3 Data

The data consists of 28 simulated directory enquiries dialogues (reported in Howes et al., 2019), with each 'caller' getting their 'operator' to look up the phone number for three business names of differing complexities (using an online phone book). As the data is simulated, the 'operators' are non-professionals, which is positive for our purposes, as each pair had to develop their own strategies for managing potential miscommunications rather than having been trained in any specific methods (such as using the "alpha", "bravo", "charlie", NATO phonetic alphabet). The majority of the participants were non-native English speakers.

4 Method

We extract sequences from the dialogues where the caller requests information regarding the business name, with each sequence beginning when the name (or part thereof) is first mentioned (line 9-10 in example 1), and the final step being when the operator demonstrates success by finding the phone number (line 32 in example 1). This process results in 84 sub-sequences from the dialogues, with these sub sequences ranging from 4 utterances (3 turns) to 119 utterances (73 turns). Note that in 3 of the 84 cases, the operator did not resolve the business name and could not locate the phone number.

4.1 Annotation tags

We make use of the manual annotations in the corpus, with the overview of annotations used shown in Table 1.[4]

[3] The complete corpus (transcriptions, audio and annotations) is available on the Open Science Framework (`osf.io/2vjkh`; Bondarenko et al., 2019).

[4] The tag names have been modified compared to (Bondarenko, 2019) for clarity.

Tag	Value	Explanation
`acknowledge (Ack)`	y/n	For all utterances: does this sentence contain a backchannel (e.g. 'yeah', 'mhm', 'right') or a repeated word or phrase acknowledging the proposition or speech act of a previous utterance? (Note 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
`clarify (CL)`	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 1: Annotation Tags

4.2 Feedback subtypes annotation

We also made use of the feedback subtypes annotation from Bondarenko (2019). For acknowledgements these are:

`Ack(cont)` continuers, i.e. acknowledgement/backchannel words like "okay", "yeah", "yes", "mmhm" (e.g. Example 1; line 8).

`Ack(verb)` verbatim repetitions of (parts of) previous utterances (e.g. Example 1; line 27)

For clarification requests these are:[5]

`CR(gen)` – General request, indicates a non-specific lack of perception/understanding of other speaker's previous utterance (e.g. "sorry?", "what?")

`CR(rep)` – Repetition request, asks other speaker to repeat a previous utterance (e.g. Example 1; line 11)

`CR(conf)` – Confirmation request, asks other speaker to provide a confirmation (e.g. Example 1; line 17)

`CR(spell)` – Spelling request, asks other speaker to spell out the name of the queried business or its address (e.g. "could you spell that for me please?", "is that a W?")

4.3 Content annotation

As we are focusing on the business name sequences, and how interlocutors handle unfamiliar names with possibly idiosyncratic spelling, we also annotated for how spelling sequences are initiated and carried out.

`Name` speaker mentions the name of a business or its address in full or in part

`Spell` speaker provides a spelling for the name or the address of a business in full or in part, usually in instalments of one or more letters

`Spell-offer` speaker offers to provide a spelling

`Spell-accept` speaker accepts a `Spell-offer`

`Spell-reject` speaker rejects a `Spell-offer`

5 Finite-state model of grounding names

As the next step in our analysis we present a tentative formalisation of the interactive process of grounding of names following Traum's descriptive finite-state model of grounding in dialogue (Traum, 1994). This is intended as a first step towards a full computational model, which will include information state updates corresponding to conversational moves based on the model of incremental processing proposed by Skantze and Schlangen (2009).

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. The model was created by analysing the relevant sequences in the corpus and investigating recurrent patterns of feedback. 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 9 possible states with states S and F representing, respectively, a state where an utterance containing a name has not been initiated

[5] The categories for acknowledgements may conflate form and function, whilst those for CRs do not consider the form. This may mean that we miss important parallels or differences between acknowledgements and clarification requests and we intend to address this in future work.

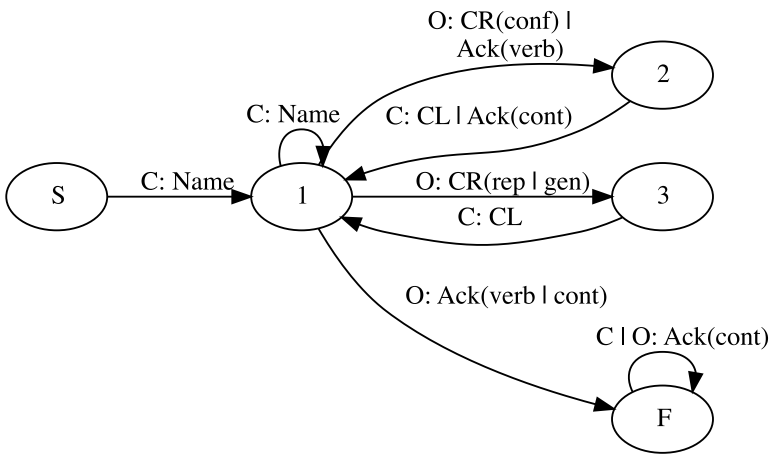

Figure 1: Finite state model of name phase

yet (S), and a state where it has been grounded (F). For readability, the network has been split into two parts: the initial phase where the name is uttered (and possibly clarified and/or acknowledged) (Figure 1) and the (optional) spelling phase (Figure 2).

5.1 Name phase

Figure 1 shows the finite state model of the Name phase. The transition from state S to state 1 corresponds to the first mention of the name or part of the name by the caller (S→1). In the simplest case, the initial name or name part (or sequence of names or name parts) is followed by an acknowledgement by O (1→F), as in (2).[6]

(2) DEC11:8-9

8	C	(. . .) it's called the good earth	Name	S→1
9	O	yes let me look that up real quick	Ack(cont)	1→F

The initial name can sometimes be followed by a self-repetition or a continuation of the name, hence there is a possibility for recursion in this state (1 →1).

In response, O can issue an Ack(verb) (1→2). C can respond by acknowledging, as in (3), or by providing the name again, possibly correcting (for example if O's hypothesis was wrong) the name (as in 4).

(3) DEC01:106–109

106	C	yeah er the name of the place it's sweet things	Name	S→1
107	O	sweet things	Ack(verb)	1→2
108	C	yeah exactly	Ack(cont)	2→1
109	O	okay	Ack(cont)	1→F

(4) DEC01:64-66

64	C	uh er the place is er todich	Name	S→1
65	O	todit uh i	Ack(verb)	1→2
66	C	uh todich (. . .)	CL	2→1

O can also provide one or more CR(conf) followed by CL from C:

(5) DEC07:62-68

62	C	and er the next number i'm looking for is the peasant	Name	S→1
63	O	the peasant?	CR(conf)	1→2
64	C	yes	CL	2→1
65	O	like in uh farmer?	CR(conf)	1→2
66	C	exactly	CL	2→1
67	O	the peasant	Ack(verb)	1→F
68	O	okay	Ack(cont)	F→F

Alternatively, O can ask for a repetition ("Please repeat that.") or provide generic feedback ("Sorry?") (1→3) which is responded to using a (repetition of) the name (3→1), as in example 1 (relevant subsection repeated here as 6).

[6] One may want to collapse Ack(cont) in response to Ack(verb) with CL in response to CR(conf), as they can both be seen as addressing a grounding question which is raised implicitly by Ack(verb) but explicitly by CR(conf).

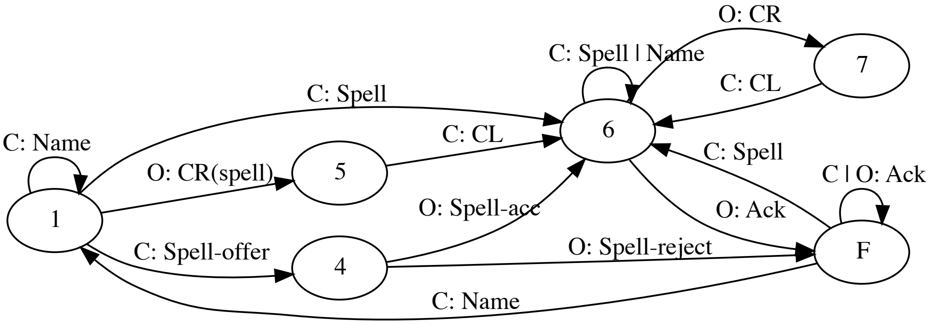

Figure 2: Finite state model of spelling phase

(6) DEC07:62-68

9	C	and the first		
10		one is rowans tenpin bowl	Name	S→1
11	O	can you repeat that for me?	CR(rep)	1→3
12	C	rowans tenpin bowl	Name	3→1

5.2 Spelling phase

The finite state network for the spelling phase is shown in Figure 2. Note that states 1 and F are the same as in Figure 1, so that the two figures show parts of a single network.

There are three ways of initating spelling, the most straightforward being where C simply starts spelling after the initial name (1→6):

(7) DEC01:66-78

66	C	uh todich is like er	Name	2→1
		T for er er thailand	Spell	1→6

As an aside, we may note from (7) that the proposed model does not currently distinguish different ways of spelling (alphabetically, using the NATO phonetic alphabet, or using some other ad hoc convention such as country names – see Howes et al., 2019 for discussion of the different strategies used). A system should be able to interpret all these strategies as instances of spelling, but should perhaps respond in a consistent and normative way (e.g. using the NATO phonetic alphabet, as a trained directory enquiries operator would).

Second, O may issue a clarification request for spelling (1→5) which is (optionally) followed by a CL from the caller (5→6):

(8) DEC25:15-18

15	C	silver cross	Name	S→1
16	O	can you spell that one for me please?	CR(spell)	1→5
17	C	yes	CL	5→6
18	O	S	Spell	6→6

Third, an offer to spell by C (1→4) may be followed by an acceptance by O (4→6):

(9) DEC21:5–7

5	C	(. . .) the fi- first is er chesneys	Name	S→1
6	C	uh do you want me to spell it?	Spell-offer	1→4
7	O	yes please	Spell-acc	4→6

If the offer is rejected, the final state is reached (4→F), and presumably this is taken to indicate that the name is grounded as in (10).[7]

(10) DEC28:87-90

87	C	er the name of the place is the black dog	Name	S→1
88	O	[the black dog]	Ack(verb)	(1→2)
89	C	[shall i] spell it?	Spell-offer	1→4
90	O	no i think it's okay i can try i think the black dog (. . .)	Spell-reject	4→F

State 6 and F form the "spelling loop", where spelling instalments from C (6→6) are interleaved with acknowledgements from O (F→6).

(11) DEC21:8-15

8	C	uh C H	Spell	6→6
9	O	yes	Ack(cont)	6→F
10	C	E S	Spell	F→6
11	O	yeah	Ack(cont)	6→F
12	C	N E Y S	Spell	F→6
13	O	N E Y S	Ack(verb)	6→F
14	C	yes	Ack(cont)	F→F
15	O	okay	Ack(cont)	F→F

[7] Square brackets denote overlap. Here, we assume that O's contribution in line 88 is abandoned.

In the final state, C may add a further name (or name part) (F→1), as in lines 12-16 of (1), repeated below with annotations:

(12) DEC07:13-16

13	C	so it's rowan	Name	1→1
14	C	R O W A N S	Spell	1→6
15	O	yes	Ack	6→F
16	C	tenpin	Name	F→1

The fact that each acknowledgement leads to the final state reflects one way of addressing the tricky problem of figuring out whether the spelling phase has finished, or if there is more to come. As mentioned above, the name is assumed to be fully grounded once the model reaches state F. In effect, this means that the proposed model tentatively assumes that spelling is finished after each spelling instalment. The transitions from F (F→5, F→6, F→1) require revoking the assumption that the name is fully grounded. Fortunately, this can be handled by the buffer model proposed in Skantze and Schlangen (2009).

The alternative would be to do the transition to F only when there is some reason to assume that spelling is finished, e.g. that the letters spelled out so far seem a likely complete candidate spelling for the initial word. Furthermore, intonation can be used to detect the end of a spelling phase, similar to (Skantze and Schlangen, 2009). However, this does not appear to be a trivial problem and in the current model we do not assume a solution for it. Instead, the end of a spelling phase will be signalled by some conversational move other than Ack, CR(spell) or Spell in the F state.

Spelling instalments may be responded to with a clarification request (6→7) followed by a clarification (7→6):

(13) DEC4:182-186

182	C	B O N E	Spell	6→6
183	O	B O?	CR	6→7
184	C	yes	CL	7→6
185	O	B O	Ack(verb)	6→F
186	O	N E	Ack(verb)	F→F

Note that the spelling phase sometimes only covers part of the name:

(14) DEC10:59-61

59	C	it's called lyle's	Name	S→1
60	C	with a Y	Spell	1→6
61	O	lyle's	Ack(verb)	6→F

6 Limitations of the model

In case of misunderstandings, the data suggests that they are largely resolved quickly and locally. However, there are some interesting cases where they persist, and such cases tend to be a challenge to formal modelling. In Example 15, which is arguably not covered by our current model, 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, and/or where one or both speakers are using a non-native language.

(15) DEC22:56–138

56	C	er the next one is er tanfield chambers
57	O	santias?
58	C	tanfield like t- T A N
59	O	sorry i don't hear you again please?
60	C	er T A N
61	O	C?
:	:	:
77	C	tanfield T like thomas
78		anna nora
79	O	thomas ar- okay nora okay
80		tan
81	C	er tan er
82		with a - filip with an F
83	O	filip
84		yeah
:	:	:
107	C	er
108	O	pilip
109	C	fanny
110	O	mmhm
111	C	fanny
112		ivonne
113	O	P
114		as in panda
115		right?
116	C	sorry i didn't hear you
117	O	P
118		the next one is a P
119		as in panda
120	C	P?
121		or okay
122		then
123	C	no

124		it's er
:	:	:
133	C	uh fanny
134	O	<unclear> I don't know that name funny?
135	C	yeah or like filip but with an F
136		or if you say fruits
137	O	with an F?
138		okay
139	C	F yeah

7 Conclusions and future work

Using genuine examples from a freely available directory enquiries corpus, we have provided a finite state model that encapsulates a basic component of an incremental computational account of grounding of unfamiliar names. The data also show that models which treat the word as the minimally grounded unit will fail at this task.

Formalising the model in this way reveals some potential issues with the original annotation scheme from Bondarenko (2019), such as the possible conflation of form and function in the subcategorisation of CRs and acknowledgements and suggests some ways in which the annotation scheme could be improved.

An important extension to the work presented here would be to fit the model to all of the 84 name sequences in the corpus, and see if it generalises to the number sequences (which are assumed to be simpler, but still require the types of instalments seen in the names sequences). This would enable us to get a better idea of not just the coverage of the model, but also which transitions are more common in the corpus, offering a principled way to develop a probabilistic model. Going further in this direction, we would also like to train a probabilistic FSA on a training section of the corpus and test it on unseen data.

We hope to extend the model with an information state update model for incremental grounding of names building on Schlangen and Skantze (2011).

Acknowledgements

This work was supported by two grants from the Swedish Research Council (VR): 2016-0116 – Incremental Reasoning in Dialogue (IncReD) and 2014-39 for the establishment of the Centre for Linguistic Theory and Studies in Probability (CLASP) at the University of Gothenburg.

References

Adrian Bangerter, Herbert H Clark, and Anna R Katz. 2004. Navigating joint projects in telephone conversations. *Discourse Processes*, 37(1):1–23.

Janet Beavin Bavelas, Peter De Jong, Harry Korman, and Sara Smock Jordan. 2012. Beyond backchannels: A three-step model of grounding in face-to-face dialogue. In *Proceedings of Interdisciplinary Workshop on Feedback Behaviors in Dialog*.

Anastasia Bondarenko. 2019. Grounding of names in directory enquiries dialogue. a corpus study of listener feedback behaviour.

Anastasia Bondarenko, Christine Howes, and Staffan Larsson. 2019. Directory enquiries corpus. Available at osf.io/2vjkh.

Elizabeth A Boyle, Anne H Anderson, and Alison Newlands. 1994. The effects of visibility on dialogue and performance in a cooperative problem solving task. *Language and speech*, 37(1):1–20.

Okko Buß, Timo Baumann, and David Schlangen. 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*, pages 233–236. Association for Computational Linguistics.

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

Herbert H. Clark. 1996. *Using Language*. Cambridge University Press.

Herbert H. Clark and Susan A. Brennan. 1991. *Grounding in communication*, pages 127–149. Washington: APA Books.

Herbert H. Clark and Edward A. Schaefer. 1989. Contributing to discourse. *Cognitive Science*, 13:259–294.

Herbert H Clark and Edward F Schaefer. 1987. Collaborating on contributions to conversations. *Language and cognitive processes*, 2(1):19–41.

Marcus Colman and Patrick G. T. Healey. 2011. The distribution of repair in dialogue. In *Proceedings of the 33rd Annual Meeting of the Cognitive Science Society*, pages 1563–1568, Boston, MA.

Christine Howes, Anastasia Bondarenko, and Staffan Larsson. 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.

Christine Howes and Arash Eshghi. 2017. Feedback relevance spaces: The organisation of increments in conversation. In *Proceedings of the 12th International Conference on Computational Semantics (IWCS 2017)*. Association for Computational Linguisitics.

Jacqueline C Kowtko and Patti J Price. 1989. Data collection and analysis in the air travel planning domain. In *Proceedings of the workshop on Speech and Natural Language*, pages 119–125. Association for Computational Linguistics.

Staffan Larsson. 2002. *Issue-based Dialogue Management*. Ph.D. thesis, Göteborg University. Also published as Gothenburg Monographs in Linguistics 21.

Matthew Purver. 2004. *The Theory and Use of Clarification Requests in Dialogue*. Ph.D. thesis, University of London.

Christoph Rühlemann. 2007. *Conversation in Context: A Corpus-Driven Approach*. Continuum.

David Schlangen and Gabriel Skantze. 2011. A general, abstract model of incremental dialogue processing. *Dialogue and Discourse*, 2(1):83–111.

Gabriel Skantze and David Schlangen. 2009. Incremental dialogue processing in a micro-domain. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, EACL '09, pages 745–753, Stroudsburg, PA, USA. Association for Computational Linguistics.

David Traum. 1994. *A Computational Theory of Grounding in Natural Language Conversation*. Ph.D. thesis, University of Rochester.

Thomas Visser, David Traum, David DeVault, and Rieks op den Akker. 2014. A model for incremental grounding in spoken dialogue systems. *Journal on Multimodal User Interfaces*, 8(1):61–73.