Limits of Simple Dialogue Acts for Tactical Questioning Dialogues

Ron Artstein and **Michael Rushforth*** and **Sudeep Gandhe** and **David Traum**
Institute for Creative Technologies, University of Southern California
12015 Waterfront Drive, Playa Vista, CA 90094-2536, USA

MAJ Aram Donigian
United States Military Academy
West Point, NY 10996, USA

Abstract

A set of dialogue acts, generated automatically by applying a dialogue act scheme to a domain representation designed for easy scenario authoring, covers approximately 72%–76% of user utterances spoken in live interaction with a tactical questioning simulation trainer. The domain is represented as facts of the form ⟨object, attribute, value⟩ and conversational actions of the form ⟨character, action⟩. User utterances from the corpus that fall outside the scope of the scheme include questions about temporal relations, relations between facts and relations between objects, questions about reason and evidence, assertions by the user, conditional offers, attempts to set the topic of conversation, and compound utterances. These utterance types constitute the limits of the simple dialogue act scheme.

Introduction

In previous work, we presented a spoken dialogue system for tactical questioning simulation which uses a simple scheme of dialogue acts, designed to facilitate authoring by domain experts with little experience with dialogue systems (Gandhe et al. 2009). The dialogue acts are generated automatically from a representation of facts as ⟨object, attribute, value⟩ triples and actions as ⟨character, action⟩ pairs. We found that initially the dialogue act scheme only covered about 50% of the user utterances, but our analysis showed that simple extensions could increase coverage to above 80% (Artstein et al. 2009). This paper puts that claim to test. We took a corpus of user utterances collected in interaction with the system, and mapped it to a set of dialogue acts in two stages: first we mapped half of the utterances to the original dialogue acts used in collecting the corpus, then we added facts to the domain representation in order to address gaps found in the coverage, and afterwards we mapped the held out data to dialogue acts derived from the expanded domain. The conclusion from this process is that the claim of Artstein et al. (2009) was about right – the expanded domain covers about 72–76% of the user utterances. While many of the remaining utterances could also be represented through an additional expansion of the domain, there remains a set of utterances which cannot be represented using the simple scheme. This paper presents a detailed analysis of those utterances that cannot be expected to be handled by the scheme, exploring the limits of this simple dialogue act representation.

Dialogue acts are often used as representations of the meaning of utterances in dialogue, both for detailed analyses of the semantics of human dialogue (e.g., Sinclair and Coulthard 1975; Allwood 1980; Bunt 1999) and for the inputs and outputs of dialogue reasoning in dialogue systems (e.g., Traum and Larsson 2003; Walker, Passonneau, and Boland 2001). There are many different taxonomies of dialogue acts, representing different requirements of the taxonomizer, both the kinds of meaning that is represented and used, as well as specifics of the dialogues and domain of interest (Traum 2000). There are often trade-offs made between detailed coverage and completeness, simplicity for design of domains, and reliability for both manual annotation and automated recognition. A common concern for theories of dialogue acts is representing the mechanisms that regulate the flow of conversation, which determine dialogue properties such as turn-taking, coordination among speakers and cohesiveness of the dialogue.

In our tactical questioning simulator, the scheme is intentionally kept very simple, in order to allow authoring by domain experts who work on the level of the domain representation, without detailed knowledge of dialogue act semantics and transitions (Gandhe et al. 2009). This simplicity results in limited expressibility. We found that in the specific genre of tactical questioning of a virtual character, most of the difficulties faced by the simple dialogue act scheme are not ones of regulating the conversation. Rather, it is the representation of information. The purpose of tactical questioning is to extract specific information through interview, and users consistently employ a richer view of the information than the system can represent. While the gap in coverage only affects a small fraction of user utterances, addressing it would require changes not only to the dialogue act scheme, but to the domain representation as well. This paper provides a characterization of the tactical questioning domain as it appears from an interviewer's perspective, based on an analysis of actual user utterances.

The remainder of the paper describes the tactical questioning genre of dialogue and the dialogue system architec-

*Now at the University of Texas at San Antonio

ture used in collecting the corpus; presents the corpus and the procedure for annotation and domain expansion; and presents the results of the annotation experiment, both in quantitative terms (reliability and coverage) as well as a detailed analysis of the gaps of the dialogue act representation.

Tactical Questioning

Artstein et al. (2009) provides an overview of the Tactical Questioning domain, which is defined as "the expedient, initial questioning of individuals to obtain information of immediate value" (U.S. Army 2006). A tactical questioning dialogue system is a simulation training environment where virtual characters play the role of a person being questioned; these characters display a range of behaviors such as answering questions cooperatively, refusing to answer questions, or intentionally providing incorrect answers (lying). The interviewer (human participant) may work to induce cooperation by building rapport with the character, addressing their concerns, making promises and offers, as well as threatening or intimidating the character.

System architecture

The architecture for our tactical questioning dialogue systems is a compromise between a text-to-text classifier that directly maps questions to responses in a stateless fashion (Leuski et al. 2006) and a full-fledged system with intricate reasoning and inference capabilities (Traum et al. 2008). It employs a fairly basic representation of dialogue acts, which are generated automatically from a simple domain representation. The generated dialogue acts reflect the role of the human participant as an interviewer and the character as a person being interviewed. We thus make a distinction between *user dialogue acts* and *character dialogue acts* – some dialogue act types are made by both user and character, but others are restricted to only one of the participants.

The dialogue acts are employed in conversation through a finite-state representation of local dialogue segments, a set of policies for engaging in the network, and a rule-based dialogue manager to update the context and choose dialogue acts to perform (Gandhe et al. 2008). This functionality allows for short subdialogues where the character can ask for and receive certain assurances (such as protection or confidentiality) and still remember the original question asked by the trainee. The link between dialogue acts and natural language is provided by a statistical classifier (Leuski and Traum 2008).

The domain representation encodes the character's knowledge as a set of facts of the form ⟨object, attribute, value⟩; in addition, the domain specifies a number of actions that the character and interviewer may perform, such as offers, threats, compliments and insults. Dialogue acts are automatically generated from the domain specification, by applying an illocutionary force (or *dialogue act type*) to a semantic content containing the relevant portion of the domain specification. For example, each fact generates 3 dialogue acts – a character dialogue act of type *assert*, a user dialogue act of type *yes/no question*, and a user dialogue act of type *wh-question* which is formed by abstracting over the value. Each object in the domain is considered a topic of conversation, and generates a set of grounding acts used for confirming the topic (*repeat-back* and *request-repair*). Additional dialogue act types include forward function (elicitation) and backward function (response) dialogue acts, as well as some generic dialogue acts that are defined independently of the domain such as greetings, closings, thanks, and special dialogue acts that are designed to handle out-of-domain dialogue acts from the user.

The system architecture was designed to facilitate rapid creation of characters by scenario designers who are experts in tactical questioning, but not experts in dialogue or dialogue systems (Gandhe et al. 2009). The architecture therefore hides much of the dialogue logic from the scenario designer, exposing only the domain and a limited set of policies. The simple structure of the domain representation is intended to provide a minimal amount of structure that would allow automatic creation of dialogue acts, while keeping authoring possible without extensive knowledge of ontologies. The representation is intended to capture just enough information about a user's actual utterance to allow for natural and believable dialogue behavior by the character.

Of course, users are not aware of the system's limited representations, and their models of the domain are richer than what is encoded. In a pilot study (Artstein et al. 2009) we found that the available dialogue acts adequately represented about 50% of the user utterances, and our analysis showed that with some modifications, coverage was expected to increase to 80% or above. The remaining ($< 20\%$) utterances could be dealt with using policies for unrecognized input (such as clarification requests or character initiative), which would result in a believable user experience that is useful for training.

This paper tests the claim of Artstein et al. (2009) using a corpus collected in live interaction with a virtual human with an expanded domain and dialogue act set. We found that coverage has indeed increased to 72–76%. However, there remains a substantial number of user utterances which cannot be represented using the dialogue act scheme, and we provide a detailed characterization of these utterance types.

Scenario details

The experiment reported in this paper used one specific scenario implemented in the dialogue system described above. This is the same scenario described in Artstein et al. (2009), with small modifications based on the results of that experiment. In this scenario, the user plays the role of a commander of a small military unit in Iraq whose unit had been attacked by sniper fire while on patrol near a shop owned by a person named Assad. The user interviews a character named Amani who lives near the shop, was a witness to the incident, and is thought to have some information about the identity of the attackers.

Amani's knowledge about the incident is represented as facts in the domain – triples of the form ⟨object, attribute, value⟩; each fact is either true or false (false facts are used by Amani when she wants to tell a lie). Table 1 gives some facts about the incident. For example, Amani knows that the name of the suspected sniper is Saif, and that he lives

Object	Attribute	Value	T/F
strange-man	name	saif	true
strange-man	name	unknown	false
strange-man	location	store	true
brother	name	mohammed	true

Table 1: Some facts about the incident

in the store. She can lie and say that she doesn't know the suspect's name. She does not have an available lie about the suspect's location (though she can always refuse to answer a question). The facts in the domain give rise to dialogue acts – for example, the fact ⟨strange-man, name, saif⟩ defines a character dialogue act with a meaning equivalent to "the suspect is named Saif" (*assert*), and two user dialogue acts, equivalent in meaning to "is the suspect named Saif?" (*yes/no question*) and "what is the suspect's name?" (*wh-question*).

Since our experiment is intended to check how well the dialogue act scheme represents user utterances, the remainder of the paper will be concerned only with the user dialogue acts generated by the scheme, not with the character dialogue acts or dialogue policies.

Method

We ran a pilot study at ICT, the results of which were reported in Artstein et al. (2009). Based on the pilot study we modified the domain, adding a few facts. We also made some changes to the dialogue act scheme, adding several dialogue act types that are generated from the domain. The character's policies were updated to handle the new dialogue act types.

Corpus collection

We collected a corpus of dialogues between human participants and the Amani character at the United States Military Academy at West Point. The dialogue participants were all cadets enrolled in a negotiation course; they had practiced negotiations in human-human role plays, but had never talked to a virtual character. Dialogue participants were given an instruction sheet with some information about the incident, the character, and suggestions for interaction, but no guidance about particular language to use with the character (see appendix). The character's behavior could be set to either confirm offers and topic shifts explicitly (high grounding) or not confirm them (low grounding). Each participant talked to the character twice (one interaction of each type), with the order of presentation balanced across participants; participants were not informed of the variation, and were instructed to treat the second dialogue as completely separate from the first. Since the current experiment focuses only on the user utterances and not the character behavior, we treat utterances from both conditions as a single corpus. The corpus consists of 68 dialogues (34 participants), comprising of a total of 1854 utterances; dialogue lengths vary from 8 to 46 utterances (mean 27.3, median 28.5, standard deviation 8.5).

Dialogue act annotation

Utterances were matched to fully specified user dialogue acts by 3 experienced annotators, including the first and second authors and a student annotator. The annotation guidelines were to match each user utterance to the most appropriate user dialogue act, and if no dialogue act was close enough, to match to "unknown". Based on the problems reported in Artstein et al. (2009), we added instructions to treat *Do you know* and *Can you tell* questions as wh-questions, and to treat formulaic greetings such as *How are you* and *It's nice to meet you* as greetings rather than questions or assertions.

Matching utterances to dialogue acts was done in two rounds. For the first round, the corpus was split in the following fashion. Whole dialogues were randomly selected until they totaled more than 100 utterances; this portion was annotated independently by all annotators and served as a reliability sample. The remaining dialogues were randomly assigned to annotators in a way that approximately balanced the number of utterances among the annotators. The annotators then matched utterances to dialogue acts from the system employed in collecting the corpus, using the domain creation tool (Gandhe et al. 2009), until about half of the corpus was annotated (annotators worked at different rates, so the number of utterances annotated at this stage was not balanced; see Table 5 below). The resulting annotated corpus will be referred to as the *original domain*, and it contains 768 unique utterances. Due to technical limitations, annotators mapped each utterance text to a single dialogue act, not taking into account context that would disambiguate different dialogue acts for the same text appearing at different times.

Based on the annotation of the original domain, we expanded Amani's domain to include meaning representations for most of the user questions that were not successfully mapped to dialogue acts. This resulted in a doubling of the number of available dialogue acts for interpretation (Table 2). The bulk of the expansion occurred in the representation of user questions through the addition of domain knowledge: each addition of a full ⟨object, attribute, value⟩ triple generated a *wh-question* and a *yes/no question*, while an addition of ⟨object, attribute⟩ without a value generated only a *wh-question* (the latter are questions that Amani can understand but does not know an answer to; such tuples were added in order to expand coverage of the user questions without adding knowledge to the character). In the course of adding domain knowledge, six new objects were created, and thus there were corresponding increases in grounding dialogue acts (*repeat-back* and *request-repair*). The *response* category includes responses to certain acts such as compliments, apologies and thanks; the increase in responses comes from the addition of compliments by Amani. No changes were made to the dialogue act scheme, that is to the rules that generate individual dialogue acts from the domain.

After expanding the domain, we took the remaining (unannotated) utterances and split them among the annotators using a similar method to the first round, creating a reliability sample of just over 100 utterances and splitting the

Dialogue Act Type	Pilot	Original	Expanded
generic acts[a]	10	13	13
closing	3	3	3
compliment	3	1	2
insult	2	2	2
offer	3	3	3
pre_closing	3	3	3
repeat_back	10	9	15
request_repair_attribute		9	15
request_repair_object	10	9	15
response	3	6	12
wh-question	**31**	**42**	**119**
yes/no question	**35**	**43**	**85**
Total	113	143	287

[a]One each of accept, ack, apology, greeting, offtopic, refuse_answer, reject, request_repair, thanks, and unknown; the original and expanded domains added clarify_elicit_offer, yes, and no.

Table 2: User dialogue acts in the Amani domain

remainder evenly among the annotators. These were then annotated by the same 3 annotators from the first round, using the same tools and instructions. The resulting annotated corpus will be referred to as the *expanded domain*, and it contains 799 unique utterances.

Results

Reliability

As a means of checking that the annotators had a similar understanding of the task, we calculated inter-annotator reliability using Krippendorff's α (Krippendorff 2004). Reliability is normally taken as a measure of the reproducibility of the annotation procedure, as codified in an annotation manual. In our case, however, the annotators were not working from detailed written guidelines; any shared understanding must therefore come from their previous experience. Reliability is therefore indicative of how straightforward the task is *before* implementing corrective measures such as detailed guidelines and domain and dialogue act improvements.

In addition to calculating agreement on the actual annotation (fully specified dialogue acts), we calculated the implicit agreement on whether a particular utterance was covered by the domain. This implicit agreement on coverage was calculated by collapsing all of the categories other than "unknown" into a single label. Table 3 shows the results of both calculations on the reliability samples for the original domain and the extended domain; the results from the pilot of Artstein et al. (2009) are also quoted here for comparison.

For the original domain, reliability was essentially the same as in the pilot: substantially above chance, but not as high as typically accepted norms. For the expanded domain we see a marked improvement in reliability, which indicates that the task is easier. The annotators and the guidelines were the same for both the original domain and expanded

	N	Individual acts			Implicit coverage		
		α	$A_o{}^{(a)}$	$A_e{}^{(a)}$	α	$A_o{}^{(a)}$	$A_e{}^{(a)}$
Pilot	224	0.49	0.55	0.11	0.38	0.74	0.58
Original	90[b]	0.49	0.58	0.19	0.33	0.67	0.52
Expanded	110[b]	0.63	0.65	0.07	0.39	0.79	0.66

[a]Krippendorff's α is defined in terms of observed and expected disagreement: $\alpha = 1 - D_o/D_e$. For expository purposes we have converted these into values representing observed and expected agreement: $A_o = 1 - D_o$, $A_e = 1 - D_e$.

[b]Several items were excluded from the reliability sample because they were not marked by all annotators.

Table 3: Inter-annotator reliability

domain, so the improvement in reliability is probably attributable to the better coverage of the domain.

The improvement in the reliability of matching utterances to specific dialogue acts does not carry over to the decision of whether an utterance is covered by the domain: here, the observed agreement of the expanded domain has gone up but so has the expected agreement, and consequently the reliability is at about the same level as the original domain. Our interpretation is that this remains a difficult decision for human judges – while domain coverage may increase, the boundary between what is covered and what is not remains fuzzy.

As an example of the fuzziness of the boundary we can take a fairly common follow-up on Amani's assertion that the suspect regularly has tea with the shopkeeper.

> Uh when he was having tea, was it close to where we are right now?
>
> Who was he having tea with?

While many such questions were judged to be out of domain, there was disagreement regarding the above two questions (and several others), on whether they were truly out of domain or if they could be mapped to questions about the suspect's location or daily routine, respectively. The expanded domain added several facts about the suspect's tea partner and drinking routine, so the above questions fall squarely within the expanded domain. However, expanding the domain did not make the domain's boundary any clearer: annotators disagreed on whether the following question could be mapped to a general question about the tea partner, or if it was outside the expanded domain.

> Why do you think he was having tea with the set?

We see that while adding facts to the domain increases the character's knowledge and thus its ability to understand user utterances, it does not necessarily make the boundaries of the character's knowledge any clearer.

Similar conclusions come from looking directly at the classification of the utterances in the reliability sample. Table 4 shows how many utterances in the reliability sample were mapped to a specific act as opposed to being judged to be out of domain, and whether the annotators agreed or disagreed about the mapping. In both the original and ex-

	Domain:	Original		Expanded	
		N	%	N	%
Specific act	Agree	32	30	53	45
	Disagree[a]	10	9	20	17
Out of domain	Agree	19	18	9	8
	Disagree[b]	46	43	35	30

[a]Utterances mapped to specific dialogue acts by all coders, where at least two coders disagreed on the dialogue act.

[b]Utterances mapped to specific dialogue acts by some coders and to "unknown" by other coders.

Table 4: Agreement on dialogue acts

Annotator	Original domain			Expanded domain		
	Total	In-domain		Total	In-domain	
		N	%		N	%
All	768	477–523	62–68	799	572–607	72–76
A	185	150	81	308	242	79
B	492	292	59	362	310	86
C	288	176	61	356	217	61

Table 5: Domain coverage

panded domain studies, the majority of disagreements are not on which dialogue act an utterance should be mapped to, but rather on whether an utterance is close enough to an existing dialogue act. The proportion of utterances mapped to specific dialogue acts is greater in the expanded domain, but the proportion of utterances on which there is agreement has not improved by much.

Domain coverage

We can define the overall coverage of a domain as the proportion of user utterances that are mapped to specific dialogue acts rather than "unknown" (we define coverage in terms of unique utterance types without regard to their frequency). Table 5 shows the coverage of the original and expanded domains, broken down by annotator; the overall coverage is reported as a range because sometimes annotators disagree as to whether an utterance is covered by the domain: the lower value considers such disagreements to be out of domain, while the higher value considers them to be in domain. The table shows that expanding the domain has improved the coverage by about 10 percentage points. We also see that annotators differ in their propensity to consider utterances to be in-domain, and that this propensity varies across the samples: the improvement in the overall coverage can be attributed to one specific annotator (coder B) for whom coverage increased substantially, coupled with the fact that the utterances in the expanded domain were more evenly balanced across the three coders.[1]

[1]The person who carried out the domain expansion was coder C, who turned out to be the one least likely to map an utterance to a specific dialogue act, both before and after the expansion.

Overall, we see that domain coverage is in line with the assessment of Artstein et al. (2009), that suitable domain expansion can bring coverage to about 80% of user utterances. Of the utterances that fall outside the expanded domain, many can still be represented using the dialogue act scheme – these constitute the "long tail" of user questions which have not been encountered or anticipated by the domain creators. Among the 227 utterances classified as outside the expanded domain by at least one annotator, we identified 94 (41%) that can plausibly be used to further expand the domain (among utterances classified as out-of-domain by all annotators the proportion is 79/192, also 41%). However, there are several types of user utterances which cannot be given a suitable representation in the scheme. These utterances demonstrate the limits for the simple dialogue act representation used in our tactical questioning system.

Temporal relations A fairly common utterance type encountered in our corpus is a question relating events in time (26 of the 227 out-of-domain utterances, or 11%).

Is Assad in the shop right now?

When have you seen the sniper on the second floor?

Did you see where he went after he had tea?

Questions with a temporal component are probably motivated by the particular scenario, where the task is to find information about a person related to a particular event. However, the representation language of facts as ⟨object, attribute, value⟩ triples does not explicitly encode time. While it is possible to represent certain static temporal facts using this scheme, for example ⟨assad, time-in-shop, now⟩, extensions would be required in order to represent temporal relations between events or perform temporal reasoning. Such an extension could be, for example, adding a temporal index to each fact, though this would increase authoring complexity.

Requests for elaboration Questioners often followed up on the character's responses by asking for additional details. Often such questions ask about facts that can be represented in the scheme; some questions, however, ask explicitly about information in relation to facts that were just provided (17 of 227 utterances, or 7%).

Do you know if there are anyone else in that building?

Have you seen him anywhere else?

OK then, do you think there is another door in the shop?

The representation language derives question dialogue acts from facts consisting of ⟨object, attribute, value⟩ triples; the only relations between facts are those that occur implicitly, when two facts share an object and attribute but differ on value, or share an object but differ on attribute. For example, if the domain representation includes facts of the form ⟨building, occupant, strange-man⟩, ⟨building, occupant, . . . ⟩ then the dialogue manager can interpret the question *Do you know if there are anyone else in that building?* as asking for values that have not yet been provided. Asking for elaboration on objects and attributes while keeping the attribute

or value fixed would require moving from a hierarchical domain representation to a relational one.

Relations between objects A small number of question concern relations between objects (3 of 227 utterances, or 1%).

Could they be found in the same area as him?

Since the domain represents all facts as ⟨object, attribute, value⟩ tuples, any fact about two objects needs to be encoded by specifying one object as a dependent value of the other. Representing relationships between the two domain objects would require a move toward a relational semantics, much like the requests for elaboration above.

Reason and evidence A common type of question is to ask the character about the reasons or evidence for her assertions (19 of 227 utterances, or 8%).

Do you know why he was having tea?

How do you know this?

And did you see him actually pull the trigger

In the current domain representation, facts do not carry any additional information beyond the content of the fact itself. Adding reasons would require an extension of the representation, for example by enriching facts beyond ⟨object, attribute, value⟩, or alternatively by enabling relations between facts.

Assertions Our dialogue act model is geared towards the user questioning the character: each fact in the domain gives rise to question-type user dialogue acts, and assertion-type acts by the character. However, we do find that the users occasionally make assertions (21 of 227 utterances, or 9%).

I have a soldier who was wounded by a sniper.

My men are outside right now and we will be in this area for a long time.

Well, I noticed that you're a school teacher ma'am.

The underlying domain representation is symmetrical, so it is possible to add these facts to the user's domain, which would give rise to user dialogue acts of type *assert* and corresponding character question dialogue acts. However, the above examples show that user assertions in tactical questioning dialogues are more than mere statements of fact; having the character ask questions about these assertions would be pointless. To do something useful with these assertions, the system would require and inference component to capture the intention behind them.

Conditional offers Offers are represented in the domain by ⟨character, action⟩ pairs, where the action is a specific offer; some user offers come with conditions attached (10 of 227 utterances, or 4%).

We can discuss money if you give me more information.

If we were able to supply you with a weapon or armed protection, would you feel safe to tell us information?

Even though the instructions to the participants do not impose any penalty on making an unconditional offer such as providing safety or secrecy, it appears that the participants sometimes attach conditions to their offer as a means of leverage. Interpreting conditions for offers and designing suitable policies would require a richer representation than the current ⟨character, action⟩ form.

Topic setting A small number of utterances were attempts by the user to set the topic (4 of 227 utterances, or 2%).

Can we talk about the shooter?

I wanna talk about the sniper not guns.

The dialogue act scheme does not include moves to set the topic of conversation. This is a straightforward addition, because the system already keeps track of the conversation topic, and the scheme already includes grounding dialogue acts for confirming topics. Dialogue acts of type *set-topic* have been added to the scheme subsequent to the experiment.

Compound utterances A fair number of utterances consisted of multiple questions strung together (20 of 227 utterances, or 9%).

Ma'am how do they look like? Are they tall? Are they short? Do they have black hair or mustache?

Do you know where he was located? Was he in a building or was he in a mosque or something like that?

Since the system assigns a single dialogue act to each user speech event (delimited by a press and release of a button), these compound utterances cannot be represented. The proper way to deal with them is by adding a module that splits them into smaller units that can be interpreted.

Conclusion

Our study has shown that a set of dialogue acts, generated automatically from a domain representation designed for easy scenario authoring by domain experts with little detailed knowledge of dialogue systems, can achieve substantial coverage of actual user utterances employed in live conversation with a virtual character. After an initial domain has been adjusted and augmented based on several hundred user utterance, coverage rises to approximately 72%–76% of unseen utterances. Combined with dialogue management techniques to recover from misunderstandings, this level of coverage should be sufficient to allow a character to sustain a coherent interaction with the user.

Among those utterances that are not covered, the largest group (around 40%, or 12% of the total utterances) are utterances that do fit in the scheme, but have not been encountered or anticipated by the domain creators. It is inevitable that such a "long tail" of rare unseen utterances should exist. The remaining out-of-domain utterances, about 17% of the total, consist mostly of the following types: questions about temporal relations, relations between facts and relations between objects, questions about reason and evidence, assertions by the user, conditional offers, attempts to set the topic of conversation, and compound utterances. Most of these utterance types fall outside the representation capability of the system, and thus constitute the limits of the simple dialogue act scheme.

We end with a caveat about our results. Our corpus of user utterances has been collected using one specific scenario, which may have influenced the questions the users wanted to ask. For example, the large number of questions about temporal relations is probably due to the fact that the users are tasked with finding information related to an event. Our user group was also fairly homogeneous, consisting of military cadets enrolled in a negotiation course, which may have influenced their approach and strategies employed in the interaction. We expect that a different scenario or a different population of users may give rise to a somewhat different distribution of utterances. Nevertheless, we believe that this study is a good start for exploring how far the simple dialogue act representation can take us, and what actual user utterances lie beyond its scope.

Acknowledgments

The project or effort described here has been sponsored by the U.S. Army Research, Development, and Engineering Command (RDECOM). Statements and opinions expressed do not necessarily reflect the position or the policy of the United States Government, and no official endorsement should be inferred.

References

Allwood, J. 1980. On the analysis of communicative action. In Brenner, M., ed., *The Structure of Action*. Basil Blackwell. Also appears as Gothenburg Papers in Theoretical Linguistics 38, Dept of Linguistics, Göteborg University.

Artstein, R.; Gandhe, S.; Rushforth, M.; and Traum, D. 2009. Viability of a simple dialogue act scheme for a tactical questioning dialogue system. In *DiaHolmia 2009: Proceedings of the 13th Workshop on the Semantics and Pragmatics of Dialogue*.

Bunt, H. C. 1999. Dynamic interpretation and dialogue theory. In Taylor, M. M.; Néel, F.; ; and Bouwhuis, D. G., eds., *The Structure of Multimodal Dialogue, Volume 2*. Amsterdam: John Benjamins.

Gandhe, S.; DeVault, D.; Roque, A.; Martinovski, B.; Artstein, R.; Leuski, A.; Gerten, J.; and Traum, D. 2008. From domain specification to virtual humans: An integrated approach to authoring tactical questioning characters. In *proceedings of Interspeech 2008*.

Gandhe, S.; Whitman, N.; Traum, D.; and Artstein, R. 2009. An integrated authoring tool for tactical questioning dialogue systems. In *6th Workshop on Knowledge and Reasoning in Practical Dialogue Systems*.

Krippendorff, K. 2004. *Content Analysis: An Introduction to Its Methodology*. Thousand Oaks, California: Sage, second edition. chapter 11, 211–256.

Leuski, A., and Traum, D. 2008. A statistical approach for text processing in virtual humans. In *Proccedings of 26th Army Science Conference*.

Leuski, A.; Patel, R.; Traum, D.; and Kennedy, B. 2006. Building effective question answering characters. In *Proceedings of the 7th SIGdial Workshop on Discourse and Dialogue*, 18–27. Sydney, Australia: Association for Computational Linguistics.

Sinclair, J. M., and Coulthard, M. 1975. *Towards an Analysis of Discourse: The English Used by Teachers and Pupils.* Oxford University Press.

Traum, D. R., and Larsson, S. 2003. The information state approach to dialogue management. In van Kuppevelt, J., and Smith, R. W., eds., *Current and New Directions in Discourse and Dialogue*. Dordrecht: Kluwer. chapter 15, 325–353.

Traum, D.; Swartout, W.; Gratch, J.; and Marsella, S. 2008. A virtual human dialogue model for non-team interaction. In Dybkjær, L., and Minker, W., eds., *Recent Trends in Discourse and Dialogue*, volume 39 of *Text, Speech and Language Technology*. Dordrecht: Springer. chapter 3, 45–67.

Traum, D. 2000. 20 Questions on Dialogue Act Taxonomies. *J Semantics* 17(1):7–30.

U.S. Army. 2006. Police intelligence operations. Field Manual FM 3-19.50, U.S. Army. Appendix D: Tactical Questioning.

Walker, M. A.; Passonneau, R.; and Boland, J. E. 2001. Quantitative and qualitative evaluation of darpa communicator spoken dialogue systems. In *ACL '01: Proceedings of the 39th Annual Meeting on Association for Computational Linguistics*, 515–522. Morristown, NJ, USA: Association for Computational Linguistics.

Appendix: participant instructions

The following information sheet was given to all experiment participants, to serve as background while talking to the character.

Situation: You are a 2LT Platoon Leader, stationed in a small village in Iraq. While on patrol yesterday, your platoon came under sniper fire, which seriously wounded one of your soldiers. Local intelligence indicates a woman named Amani witnessed the sniper.

Mission: You will question Amani Omar Al-Mufti in order to determine the location and appearance, and daily activities of the sniper that wounded the soldier.

Execution: You received permission from Amani's eldest brother to question her. He is present during the questioning to act as a chaperone, however, you will not need to speak any further with the brother. Your platoon will provide security outside during your questioning inside. Gather intelligence from Amani and offer to keep her family safe if she shows concern.

If Amani becomes too hostile or indicates that she no longer has time, end the interview before too much ill will is generated, without pressing her on any issues. You may have the opportunity to meet with her in the future.

Service Support: N/A

Command and Signal: N/A

Screening Report

A: Report Number: DTG:

B: Capture Data

N/A

C: Biographical Information

Full Name/ Rank/ Service Number:
  a. Amani Omar Al Mufti
  b. Civilian
  c. N/A

Date/ Place of Birth:
  a. 16AUG1983
  b. Local

Sex/ Marital Status/ Religion:
  a. Female
  b. Single
  c. Islam (Shiite)

Full Unit Designation/ Unit Code:
  a. N/A
  b. N/A

Duty Position:
  a. Housekeeper and Guardian of Siblings
  b. Teacher at private K-12 school

Military Education/ Experience:
  a. N/A
  b. N/A

Civilian Education/ Experience:
  a. Completed Secondary School, some college
  b. She is an English teacher at a K-12 school.

Languages Spoken (Fluency):
  a. Arabic (Native)
  b. English (Fluent)

D: Observations

Physical Condition:
  a. No Issues

Uniform Type/ Condition:
  a. N/A
  b. N/A

Assessment of Knowledgeability:

She is likely to have personal knowledge about the gunman's appearances and his location.

E: Recommendations

Relationship Building:
Begin the questioning with greeting Amani. Gaining her trust and comfort is key to getting any answers from her.

Information Gathering:
Focus on finding out what she knows about the suspected sniper, his location and reasons she suspects him. If being friendly and respectful is not effective, explain to her that she and her family can have protection. If she wants anything in return for information, you are free to make an offer or refuse to make one. Make sure she understands that you value the importance of secrecy due to the sensitive nature of the visit.