The Indiana "Cooperative Remote Search Task" (CReST) Corpus

Kathleen Eberhard*, Hannele Nicholson*, Sandra Kübler†, Susan Gundersen*, Matthias Scheutz†

* University of Notre Dame
Notre Dame, IN 46556, USA
{eberhard.1,hnichol1, sgunder2}@nd.edu
† Indiana University
Bloomington, IN 47405, USA
{skuebler,mscheutz}@indiana.edu

Abstract

This paper introduces a novel corpus of natural language dialogues obtained from humans performing a *cooperative, remote, search task* (CReST) as it occurs naturally in a variety of scenarios (e.g., search and rescue missions in disaster areas). This corpus is unique in that it involves remote collaborations between two interlocutors who each have to perform tasks that require the other's assistance. In addition, one interlocutor's tasks require physical movement through an indoor environment as well as interactions with physical objects within the environment. The multi-modal corpus contains the speech signals as well as transcriptions of the dialogues, which are additionally annotated for dialog structure, disfluencies, and for constituent and dependency syntax. On the dialogue level, the corpus was annotated for separate dialogue moves, based on the classification developed by Carletta et al. (1997) for coding task-oriented dialogues. Disfluencies were annotated using the scheme developed by Lickley (1998). The syntactic annotation comprises POS annotation, Penn Treebank style constituent annotations as well as dependency annotations based on the dependencies of *pennconverter*.

1. Introduction

Coordinating natural language dialogues in cooperative naturalistic tasks where spatially separated humans need to communicate via audio links are of great utility to a variety of research programs: (1) *computational linguistics*, (2) *psycholinguistics*, and (3) *human-computer* or *human-robot interaction*. For computational linguists, they present a challenge for current parsing methods due to effects of spontaneous speech. For psycholinguists, they provide important information about cognitive processes like joint attention as well as the effects of shared knowledge and common ground on language understanding and production. For human-machine interaction designers, they provide a benchmark for natural language capabilities required for autonomous systems to interact *naturally* with humans (Scheutz et al., 2007).

This paper introduces a novel corpus of natural language dialogues obtained from humans performing a *cooperative, remote, search task* (CReST) as it occurs naturally in a variety of scenarios (e.g., search and rescue missions in disaster areas). Although there are a number of corpora of human dialogues, such as the HCRC Map Task Corpus (Anderson et al., 1991) or the SmartKom data collection (Steininger et al., 2002), this corpus is unique in that it involves remote collaborations between two interlocutors who each have to perform tasks that require the other's assistance. In addition, one interlocutor's tasks require physical movement through an indoor environment as well as interactions with physical objects within the environment. Thus, the dialogue should be particularly useful for designing human-robot interaction systems where robots have to perform physical tasks based on instructions given by a human supervisor (e.g., search and rescue missions in disaster areas).

The multi-modal corpus, CReST, contains the speech signals as well as transcriptions of the dialogues, which are additionally annotated for dialog structure, disfluencies, and for syntax. The syntactic annotation comprises POS annotation, Penn Treebank (Marcus et al., 1993) style constituent annotations as well as dependency annotations based on the dependencies of *pennconverter* (Johansson and Nugues, 2007). The corpus is available free of charge for research purposes but requires a license.

2. The Collaborative Task

2.1. Participants

Seven dyads of native American English speaking adults (between 19 and 25 years of age) participated in the experiment. The dyads consisted by chance of same sex individuals: four male dyads and three female dyads. Three male dyads and one female dyad were acquainted with each other prior to the experiment. The participants were recruited from undergraduate and graduate summer courses and paid either $5.00 or $10.00 each depending on their performance in the experiment.

2.2. Task

The experiment involved a search task that required the individuals in a dyad to coordinate their actions via remote audio communication to accomplish several tasks with target objects ("colored boxes") that were scattered throughout an indoor search environment. The environment consisted of six connected office rooms and a surrounding hallway. Neither individual was familiar with the environment before the experiment. One individual was designated as the director (D), and the other was designated as the searcher (S). The director communicated remotely with the searcher via a handsfree telephone connection in a room serving as the "home base", which was in a building near the one containing the search environment. The director was seated in front of a computer screen that displayed a map of the search environment (see Figure 1). The director was fitted with a free-head eyetracker (Applied Science

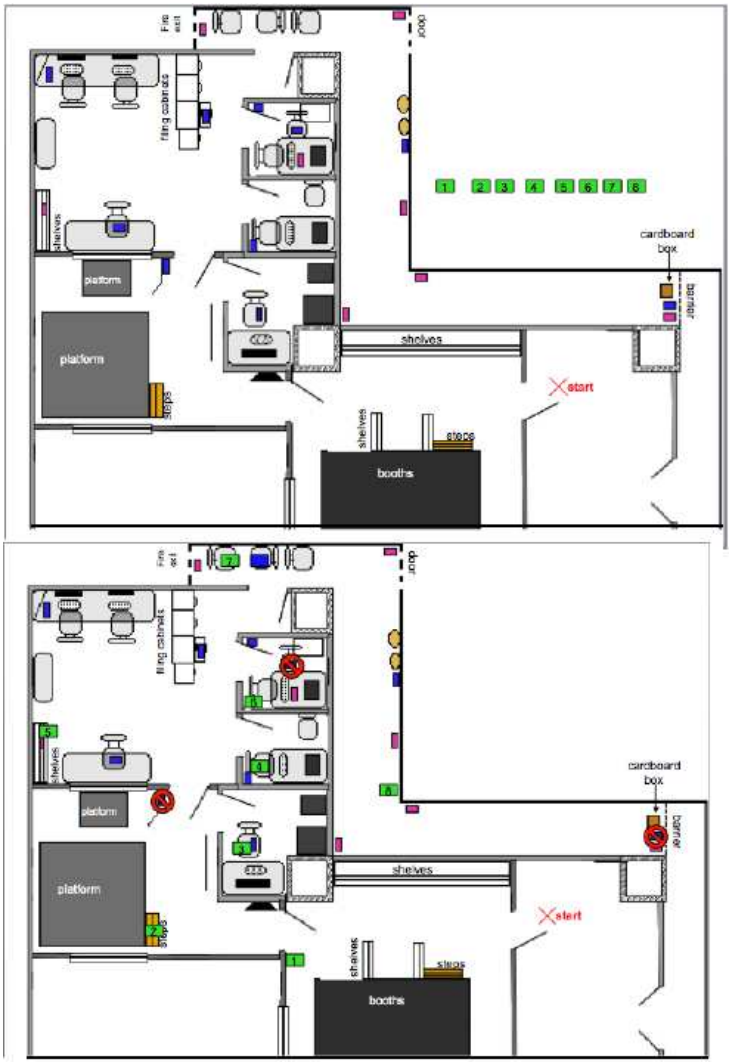

Figure 1: The top panel shows the map of the search environment that was displayed to the director on a computer screen. The bottom panel shows the actual locations of the eight green boxes in the environment, which the director was to indicate on the map by using the computer mouse to drag the boxes to the appropriate locations based on the searcher's descriptions. The three blue boxes with a circle in the bottom panel are boxes that did not exist in the environment. In addition, there was one blue box (next to green box 7) that did exist in the environment, but was not shown on the director's map.

Laboratories, Model 501) that synchronously recorded his or her eye fixations on the map along with the spoken dialogue, which was captured by a microphone positioned between the director and the telephone's speaker. The video and audio were recorded onto Hi8 video tape at a NTSC 60Hz sampling rate. The searcher wore a helmet with a cordless phone and a light-weight digital video camera that recorded his or her movement through the environment as viewed from his or her perspective and provided a second audio recording of the spoken dialogue. The digitized video was recorded in DivX4 (Perian) format with 29.97 FPS.

The experiment began with both the director and searcher at the home base, where they were given instructions for the search task. They were told that the director's map showed the locations of a cardboard box, a number of blue boxes containing colored wooden blocks, and eight empty pink boxes. They were also informed that the search environment contained eight numbered green boxes not shown on the director's map. They were told that there were several tasks that needed to be completed as quickly as possible: First, the director was to direct the searcher to the cardboard box at the furthest point in the search environment. The searcher was to collect the blocks from the blue boxes and put them into the cardboard box. This task was complicated by the discrepancies between the map and the environment (cf. Figure 1). The searcher was to report the locations of eight green boxes in the environment, and the director was to mark their locations on the map by using the computer mouse to drag numbered icons. To examine performance under time pressure, the experimenter interrupted the director after 5 minutes into the task and informed him or her that all tasks including one additional task needed to be completed within 3 remaining minutes. The additional task required the searcher to place a yellow block (obtained from a blue box) into each of the eight pink boxes. A digital clock counting down the remaining 3 minutes was dis-

played above the map. The dyads' performance was measured with respect to the number of green boxes correctly located on the director's map (up to 8 points), the number of blue boxes emptied of their blocks (up to 8 points), and the number of pink boxes containing a yellow block (up to 8 points), for a maximum possible score of 24 (the average score was 13, range 6 - 21).

2.3. Transcriptions

The video-taped recordings of the directors' eye-movements and spoken interaction with the searcher were digitized (without compression) using Apple's Final Cut Express video editing software and saved as a project file. The stereo audio tracks were exported from the project files and saved as separate stereo audio files (.aif) with a 24 kHz (24 bit) sampling rate. Each dyad's recording contained approximately 8 minutes of verbal interactions resulting in a total transcribed corpus of about 1 hour of dialogues. Orthographic transcriptions were made using Praat (Boersma and Weenink, 1996). Interval tiers were aligned with the visual waveform of the audio file. The audio track from the searcher's headcam recording was consulted when there was difficulty deciphering what the searcher said, typically as a result of overlapping speech with the director. The director's and searcher's utterances were transcribed on separate tiers with utterance boundaries corresponding to the beginnings and ends of a turn, which sometimes overlapped. The boundaries provide information concerning the duration of an utterance as well as the duration of pauses between utterances/turns.

3. Dialog Annotation

On the dialogue level, the corpus was annotated for dialogue structure and for disfluencies. These annotations are described below in more detail.

3.1. Dialogue Structure Annotation

Utterances were divided into separate dialogue moves, based on the classification developed by Carletta et al. (1997) for coding task-oriented dialogues. Their scheme views utterances as moves in a conversational game and classifies utterances into three basic move categories: *Initiation*, *Response*, and *Ready*. *Initiation* is further divided into INSTRUCT, EXPLAIN, QUERY-YN, QUERY-W, CHECK, and ALIGN. The following is an example for an align move:

D3: so wait you're supposed to do all the blue boxes and then the pink right? (ALIGN)

The category *Response* includes ACKNOWLEDGE, replies to wh-questions REPLY-WH, and "yes" or "no" replies REPLY-Y, REPLY-N. The following is an example of an ACKNOWLEDGE move consisting of a partial repetition of the previous speaker's move:

S4: so I'm at the doorway with the chairs (EXPLAIN)
D4: with the chairs okay (ACKNOWLEDGE)

Turns consisting of words such as `yes, yeah, yep, yup, right`, or `uh huh` are generally classified as ACKNOWLEDGMENT, but when they occur after a move that explicitly requests a confirmation, they are classified as REPLY-Y.

D3: there should be four filing cabinets there right? (ALIGN)
S3: yes (REPLY-Y)

D3: it looks like there's a blue box in one of the drawers (EXPLAIN)
S3: yeah (ACKNOWLEDGE)

Ready moves introduce a new initiation move. In our corpus, 37% of all instances of `okay` and 54% of the occurrences of `alright` were classified as READY moves. These instances were typically the first word of an intonational phrase that continued with either an instruct or explain move as illustrated by the following examples:

D1: well grab that blue box contents (INSTRUCT)
S1: okay (ACKNOWLEDGE)
D1: alright (READY) then uh go the other way down the hall (INSTRUCT)

S6: okay (READY) box number six is in the next little cubicle over (EXPLAIN)
D6: okay (ACKNOWLEDGE)

Two scorers independently classified the transcribed utterances with respect to their move, and disagreements were resolved by a third scorer.

3.2. Acknowledgment Function Coding

To further capture the discourse structure, the ACKNOWLEDGE moves were classified according to the functions identified in previous investigations (e.g. (Filipi and Wales, 2003; Gardner, 2001)). The labels and descriptions of the functions are as follows:

AGREE:	I agree with or understand you
CONTINUE:	I hear you, please continue
CONFIRM:	What you said is correct
CLOSE/OPEN:	Close or open a discourse segment about a subtask
3RD TURN:	Close a side sequence involving a question-answer adjacency pair

The classification of functions was based on the preceding move, the preceding move's terminal intonation, and the acknowledgment's position in a discourse segment. In particular, the CONTINUE function is a subcategory of the AGREE function. It is distinguished from the latter by following an INSTRUCT or EXPLAIN move that is an installment in a complex instruction or explanation, and the installment ended in rising or flat intonation indicating that more was to follow. The AGREE function occurred after the final installment or an INSTRUCT or EXPLAIN that ended with falling intonation. The 3RD TURN function is a subcategory of the CLOSE function. It is distinguished from the latter by occurring after a question-answer sequence. The CONFIRM function was performed by *right*, *yes* or a variant

that occurred after an EXPLAIN expressing uncertainty, as illustrated below:

D4: and there's two kind of round orange-ish objects I'm not sure what they are but there's should be a blue box next to those (EXPLAIN)
S4: yes (CONFIRM)
D4: okay (CLOSE)

Acknowledgments could perform both a CONFIRM and CONTINUE function, in which case they were labeled as CONFIRM-CONTINUE.

3.3. Disfluency Coding

The transcriptions have been annotated for disfluencies by an experienced coder using the scheme developed by Lickley (1998). Lickley's coding scheme categorized disfluencies into four main classes: *repetitions*, *substitutions*, *insertions*, and *deletions*. In a *repetition*, the words in the reparandum (i.e. the portion of the utterance to be fixed) and the repair are identical: `so two doorways and` **`you'll | you'll`** `be staring at the platform`. A *substitution* additionally contains substitutions of semantically or syntactically similar words in the repair: **`we got | you have`** `um like five pink boxes in the hallway`. An *insertion* involves the addition of one or more words prior to a word in the reparandum: **`how many box- | how many blue boxes`** `do we have?` In *deletions*, the speaker abandons an utterance in favor of a completely new one: `okay,` **`and then there's um [pause]`** `so in that roo:m you said you turned left`.
Repetitions, substitutions, and insertions have a definable reparandum and repair; deletions, however, usually do not have a definable repair. The region between the reparandum and repair, the interregnum, may contain silent pauses, filled pauses, editing terms (eg. `I mean`, `sorry`, etc.) or may contain nothing. In addition to speech repairs, filled pauses, prolongations, or elongated words and disfluent silent pauses were also coded. A silent pause is considered disfluent only if it occurs near another type of disfluency.
Disfluencies were also coded according to their position in the move, either *beginning* (B), *middle* (M) or *end* (E). In the *repetition* example above, the middle of the utterance was labeled M, while the substitution begins with a disfluency, hence it was labeled B. There were very few disfluencies that occurred at the ends of utterances, one of the few exceptions is the substitution `there should be a blue box in the ch- on the chair`. Filled pauses, silent pauses, and prolongations that occur within the interregnum of a speech repair are marked as `mid`-disfluency (MD). The `um` that occurs in the deletion example illustrates an MD disfluency.

4. Linguistic Annotation

On the linguistics level, the transcriptions were annotated with parts of speech (POS) using the Penn Treebank (Marcus et al., 1993) POS tagset (Santorini, 1990) and two different syntactic annotations: constituents based on the Penn Treebank annotations (Santorini, 1991) and dependencies

yeah	AP	you	PRP
let	VBI	're	VBP
's	PRP	gonna	VBG+TO
do	VB	find	VB
that	DDT	a	DT
yeah	UH	pink	JJ
		box	NN

Figure 2: Two examples with POS annotation.

based on the automatic dependency conversion from *penn-converter* (Johansson and Nugues, 2007). Before the transcriptions could be processed automatically, they were split into sentences based on intonation; and contractions such as `don't` or `I'm` were split following the Penn Treebank practice. All annotations had to be modified to suit the spoken language characteristics of the transcriptions. For the automatics annotation, the transcriptions were annotated using *tnt*, a trigram POS tagger (Brants, 2000), the Berkeley parser (Petrov and Klein, 2007a; Petrov and Klein, 2007b) for constituents, and MaltParser (Nivre, 2006) for dependencies. The annotations were then manually corrected by two experienced coders.
The annotations are described in more detail below.

4.1. POS Annotation

The POS annotation was based on the Penn Treebank POS tagset (Santorini, 1990). In order to accommodate the nature of the dialogues, the introduction of a small number of new POS tags was necessary. The first tags is **AP** for adverbs that serve for answering questions, such as `yes`, `no`, or `right`. This subtype of adverbs is especially important to recognize since they structure the dialogues. The next tag concerns substituting demonstratives (**DDT**) such as in `that is correct`. These demonstratives are considerably more frequent in spoken language than in the newspaper texts of the Penn Treebank, and their syntactic behavior is different from standard determiners. The last addition concerns imperatives (**VBI**), which occur very frequently given the nature of the collaborative task but, to the best of our knowledge, do not occur in the Wall Street Journal section of the Penn Treebank. The first sentence in Figure 2 shows an example of a sentence with all three new POS tags.
Another modification of the tagset concerns contractions such as in `you're gonna wanna turn to the right?`. It was decided to keep these words without splitting them. As a consequence, they were assigned combinations of tags such as **VBG+TO**. The second sentence in Figure 2 shows an example of such a contraction.

4.2. Constituent Annotation

The constituent annotation is based on the Penn Treebank annotations (Santorini, 1991). The annotation is concentrated on the surface form. For this reason, we did not annotate empty categories and traces. Since the collaborative task involved manoevering in an unknown environment, the annotation of grammatical functions concentrates on the functions subject (SBJ), locative (LOC), direction

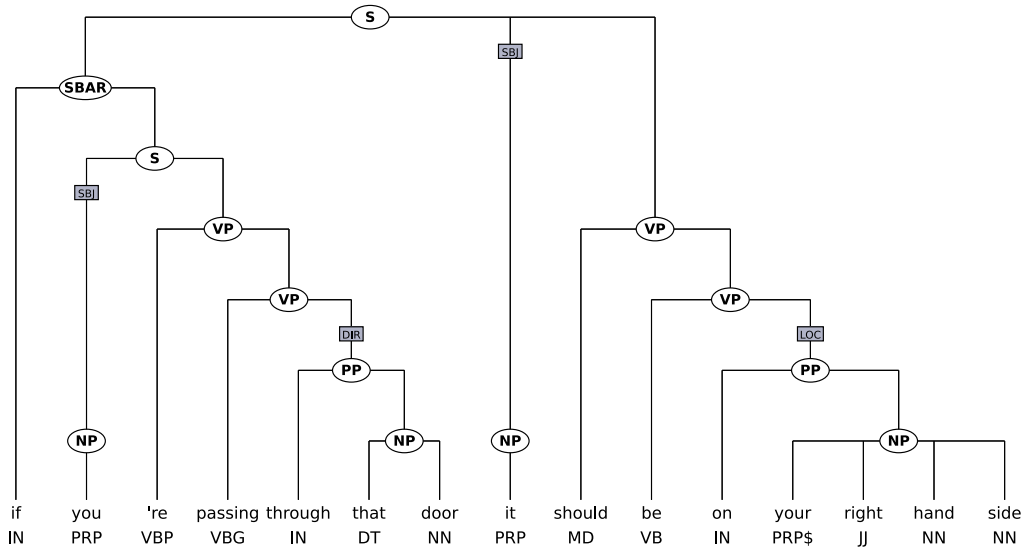

Figure 3: A constituent tree.

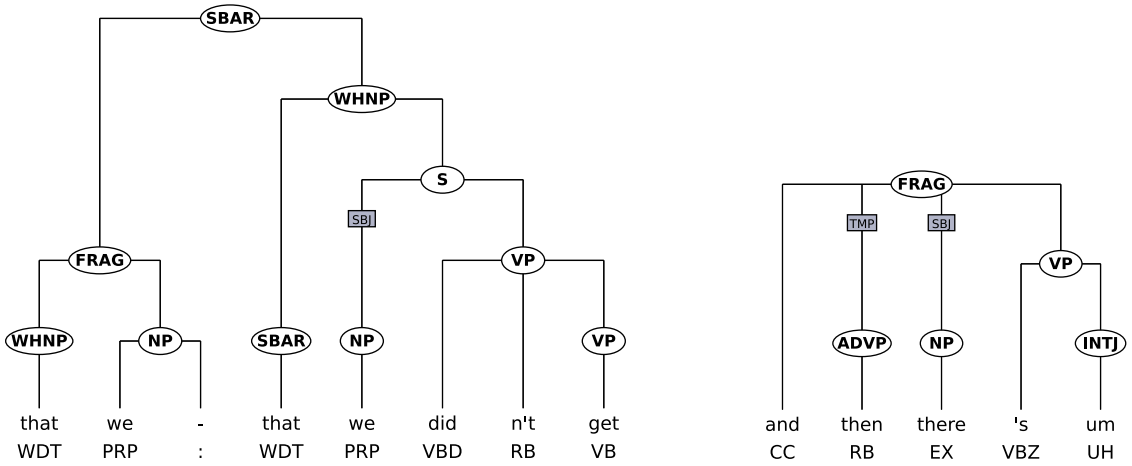

Figure 4: Constituent trees with spontaneous speech phenomena.

(DIR), and temporal (TMP). Figure 3 shows an example of the constituent annotation of a sentence that has a locative and a direction, apart from the two subjects.

Modifications of the annotation scheme were necessitated by the spontaneous speech data: For many sentences, the high frequency of disfluencies prevented a complete grammatical analysis. In such cases, the maximal possible grammatical string was annotated. The ungrammatical elements were annotated as fragments (FRAG) on the lowest level covering all the disfluencies and then integrated into the tree structure. The first utterance in Figure 4 shows an example of a repair, the second one an aborted sentence. In the first sentence, the reparandum `that we` is projected to a fragment, but attached to the whole clause. In the second sentence, the whole aborted utterance is projected to a fragment.

4.3. Dependency Annotation

The dependency annotation is based on the automatic dependency conversion from Penn-style constituents by *penn-converter* (Johansson and Nugues, 2007). This means that we used the same style of annotation, but not the converter. Instead, the sentences were parsed by a dependency parser trained on the Penn dependencies; then they were corrected manually. Figure 5 shows the sentence from Figure 3 with its dependency annotation. For coordinations, we decided to attach both the conjunction and the second conjunct to the first conjunct. The reason for this decision lies in an attempt to reach consistency with coordinations without conjunctions, for which the second conjunct would have to be dependent on the first conjunct. We also decided to make subordinating conjunctions dependent on the finite verb of the subordinate clause, which in turn is dependent on the verb of the matrix clause. For this reason, the subordinating conjunction `if` in Figure 5 depends on the verb `'re`, which depends on `should`.

Figure 5 includes a ROOT node, which is a virtual node, to which all root nodes in the sentence are attached. In the case of ungrammatical sentences, all the fragments are attached to ROOT. Ungrammatical dependencies are starred.

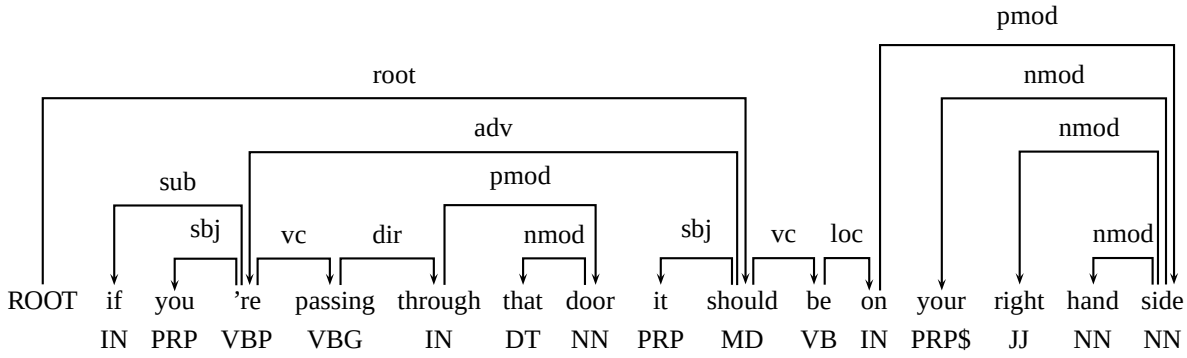

Figure 5: The dependency annotation for the sentence in Figure 3.

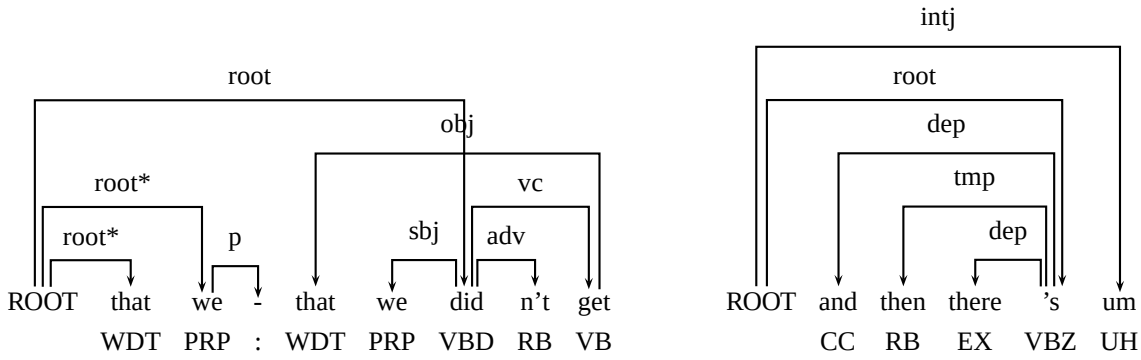

Figure 6: The dependency annotation for the ungrammatical sentence in Figure 4.

In the first sentence in Figure 6, for example, the reparandum, `that we` is dependent on the ROOT, and both dependencies are starred. This corresponds to the fragment in the constituent tree. The second sentence, in contrast, does not result in a starred dependency graph because none of the words are actually ungrammatical.

Acknowledgment

This work was in part funded by ONR MURI grant #N00014-07-1-1049 to the first and last author.

5. References

Anne Anderson, Miles Bader, Ellen Bard, Elizabeth Boyle, Gwyneth Doherty, Simon Garrod, Stephen Isard, Jacqueline Kowtko, Jan McAllister, Jim Miller, Catherine Sotillo, Henry Thompson, and Regina Weinert. 1991. The HCRC map task corpus. *Language and Speech*, 34:351–366.

Paul Boersma and David Weenink, 1996. *Praat: A System for Doing Phonetics by Computer*. Phonetic Sciences, University of Amsterdam.

Thorsten Brants. 2000. TnT–a statistical part-of-speech tagger. In *Proceedings of the 1st Conference of the North American Chapter of the Association for Computational Linguistics and the 6th Conference on Applied Natural Language Processing, ANLP/NAACL 2000*, pages 224–231, Seattle, WA.

Jean Carletta, Stephen Isard, Amy Isard, Gwyneth Doherty-Sneddon, Jacqueline Kowtko, and Anne Anderson. 1997. The reliability of a dialogue structure coding scheme. *Computational Linguistics*, 23(1):13–31.

Anna Filipi and Roger Wales. 2003. Differential uses of okay, right, and alright, and their function in signaling perspective shift or maintenance in a map task. *Semiotica*, 147:429–455.

Rod Gardner. 2001. *When Listeners Talk: Response Tokens and Listener Stance*. John Benjamins.

Richard Johansson and Pierre Nugues. 2007. Extended constituent-to-dependency conversion for English. In *Proceedings of NODALIDA 2007*, Tartu, Estonia.

Robin Lickley, 1998. *HCRC Disfluency Coding Manual*. Human Communication Research Centre, University of Edinburgh.

Mitchell Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. 1993. Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics*, 19(2):313–330.

Joakim Nivre. 2006. *Inductive Dependency Parsing*. Springer Verlag.

Slav Petrov and Dan Klein. 2007a. Improved inference for unlexicalized parsing. In *Proceedings of Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics*, pages 404–411, Rochester, NY.

Slav Petrov and Dan Klein. 2007b. Learning and inference for hierarchically split PCFGs. In *Proceedings of AAAI (Nectar Track)*, Vancouver, Canada.

Beatrice Santorini. 1990. Part-of-speech tagging guidelines for the Penn Treebank Project. Department of Computer and Information Science, University of Pennsylvania, 3rd Revision, 2nd Printing.

Beatrice Santorini. 1991. Bracketing guidelines for the Penn Treebank Project. Department of Computer and Information Science, University of Pennsylvania.

Matthias Scheutz, Paul Schermerhorn, James Kramer, and David Anderson. 2007. First steps toward natural human-like HRI. *Autonomous Robots*, 22(4):411–423, May.

Silke Steininger, Florian Schiel, Olga Dioubina, and Susen Rabold. 2002. Development of user-state conventions for the multimodal corpus in SmartKom. In *Proceedings of the 3rd Language Resources and Evaluation Conference (LREC)*, Las Palmas, Gran Canaria.