Linguistic **I**ssues in **L**anguage **T**echnology – **LiLT**
Submitted, January 2012

Parallel Syntactic Annotation in CReST

Sandra Kübler
Eric Baucom
Matthias Scheutz

Published by CSLI Publications

LiLT volume 6, issue your-assigned-Issue-Number January 2012

Parallel Syntactic Annotation in CReST

Sandra Kübler, *Indiana University* Eric Baucom, *Indiana University* Matthias Scheutz, *Tufts University*

LiLT Volume 6, Issue your-assigned-Issue-Number , January 2012.
Parallel Syntactic Annotation in CReST.

Parallel Syntactic Annotation in CReST

Sandra Kübler, *Indiana University* Eric Baucom, *Indiana University* Matthias Scheutz, *Tufts University*

November 9, 2011

Abstract

In this paper, we introduce the syntactic annotation of the CReST corpus, a corpus of natural language dialogues obtained from humans performing a cooperative, remote search task. The corpus contains the speech signals as well as transcriptions of the dialogues, which are additionally annotated for dialogue structure, disfluencies, and for syntax. The syntactic annotation comprises POS annotation, Penn Treebank style constituent annotations, dependency annotations, and combinatory categorial grammar annotations. The corpus is the first of its kind, providing parallel syntactic annotation based on three different grammar formalisms. All three annotations are manually corrected, thus providing a high quality resource for linguistic comparisons, but also for parser evaluation across frameworks.

1 Introduction

Despite the increasing interest in spoken natural language interactions in dialogue systems and with robots and other types of artificial agents, there is a suprising lack of corpora that contain typical natural language dialogue interactions in naturalistic environments. Yet, such corpora would be of great utility for developing robust components for natural processing systems for artificial agents. Specifically, they could be used to train speech recognizers and parsers, develop methods for coping with common disfluencies as they frequently occur in spontaneous speech, and define appropriate semantic formalisms that capture different non-truthfunctional aspects of typcial utterances. Moreover, they could be used as benchmarks for the systematic comparison of different speech recognizers, parsers, and semantic analyzers.

In this paper, we introduce such a corpus – the CReST corpus – which was specifically developed to fill this void. Different from standard corpora such as the Wall Street corpus of the Penn Treebank, CReST was developed with different parallel syntactic annotations in mind to specifically facilitate linguistic comparisons across grammar formalisms as well as comparison of different types of parsers (among others). As such, the corpus includes three different syntactic annotations: constituent, dependency, and combinatory categorial grammar (CCG). We start by briefly describing the corpus, followed by a description of the three types of annotation. Then we also give some natural language examples that set the CReST corpus apart from other existing corpora and point to the utility for studying and evaluation of natural language processing components in the context of naturalistic spoken language exchanges.

2 The CReST Corpus

The CReST corpus Eberhard et al. (2010) is a corpus of natural language dialogues obtained from humans performing a cooperative, remote search task in which one person outside the search environment (director) directed a person inside the environment (searcher). The director guided the searcher through the search environment, for which the director had a map, in order to find different colored boxes, enter them on the map, and place blocks in them. The director was fitted with a free-head eyetracker, and he was recorded by a microphone positioned between the director and the telephone's speaker. 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 multi-modal corpus consists of 23 dialogues. The text highlights the differences between formal written and naturally occurring language, as it is rife with directives, disfluencies, corrections, ungrammatical sentences, wrong-word substitutions, and various other constructions that are missing from written text corpora. In total, there are 40 083 words in 5 872 sentences.

The corpus contains the speech signals as well as transcriptions of the dialogues, which are additionally annotated for dialogue structure, disfluencies, and for syntax. The syntactic annotation comprises POS annotation, Penn Treebank Marcus et al. (1993) style constituent annotations, dependency annotations based on the dependencies of *penn-converter* Johansson and Nugues (2007), as well as combinatory categorial grammar annotations based on the algorithm provided by Hockenmaier and Steedman Hockenmaier and Steedman (2007).

2.1 Annotation

On the dialogue level, the corpus was annotated for dialogue structure and for disfluencies. Utterances were divided into separate dialogue moves, based on the classification developed by Carletta et al. 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 category *Response* includes ACKNOWLEDGE, replies to wh-questions REPLY-WH, and `yes` or `no` replies REPLY-Y, REPLY-N.

The POS annotation is based on the Penn Treebank POS tagset Santorini (1990), with a small number of new POS tags added to describe typical characteristics of spoken language:

- **AP** for adverbs that serve for answering questions, such as `yes`, `no`, or `right`.
- **DDT** for substituting demonstratives, such as in `that is correct`.
- **VBI** for imperatives, such as `turn left`.
- **XY** for non-words or interrupted words.

The first sentence below shows an example of a sentence with three new POS tags. Another modification of the tagset concerns informal contractions such as in `you 're gonna wanna turn to the right?`, which are kept as single words. As a consequence, they are assigned combinations of tags, such as **VBG+TO**. The second sentence below shows an example of such a contraction:

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

3 Syntactic Annotation

In addition to the levels of annotation described above, the corpus is annotated in parallel for constituent, dependency, and combinatory categorial grammar (CCG). The annotations are based on automatic annotations, either by a parser, or by conversion, and consequently manually checked. This provides a unique resource for English syntactic annotation, which allows the comparison of the different syntactic annotations for the same sentence as well as the comparison of parsers trained on the different syntactic annotations. The treebank is similar to the Turin University Treebank for Italian Bos et al. (2009), Bosco and Lombardo (2004), which covers annotations based on the same grammar formlaisms, but is more restricted in size.

3.1 Constituent Annotation

The constituent annotation is based on the Penn Treebank annotations Santorini (1991). The annotation concentrates 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), predicate (PRED), locative (LOC), direction (DIR), and temporal (TMP).

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.

3.2 Dependency Annotation

The dependency annotation is based on the automatic dependency conversion from Penn-style constituents by *pennconverter* 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. We made small changes to the annotation scheme: 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.

3.3 Combinatory Categorial Grammar Annotation

To obtain our CCG annotations, we automatically converted the Penn-style constituent annotations following the conversion by Hockenmaier and Steedman Hockenmaier and Steedman (2007) for the Penn Treebank. We then manually correct the annotations. To determine the constituent types, heuristics are required. Hockenmaier and Steedman adapted theirs from the head-finding rules developed by Collins Collins (1999) and Magerman Magerman (1994). Ungrammatical sentences are processed automatically once their constituent types are determined from the heuristics, although in such cases the terms “head,” “complement,” and “adjunct” lose some of their meaning.

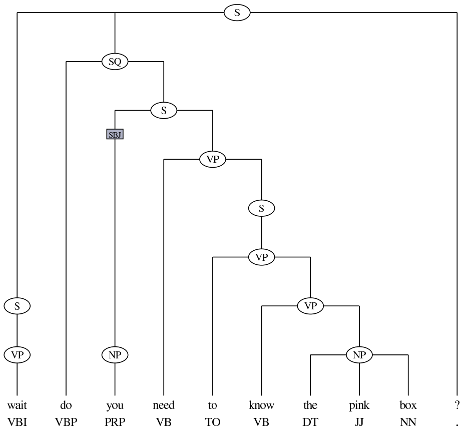
Since CReST uses additional POS tags, we added these as head candidates for `FRAG` and `VP` nodes, respectively. The heuristics used to distinguish complements and adjuncts rely on the presence of grammatical function categories, many of which are not coded in CReST. We had to disambiguate those manually. Following Hockenmaier and Steedman, we allow forward and backward rule application, and restrict the combinatory rules for CCG to forward and backward composition and backward crossing composition. This restriction sometimes leads to a proliferation in categories, especially given the fluid nature of syntax for dialogues.

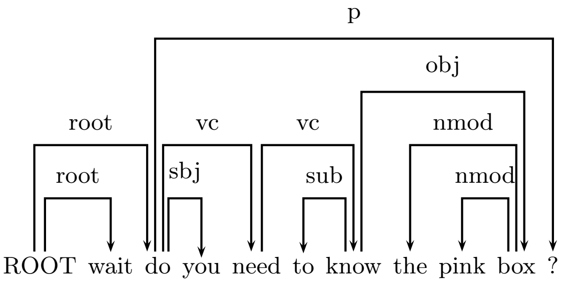
4 Selected Phenomena in CReST

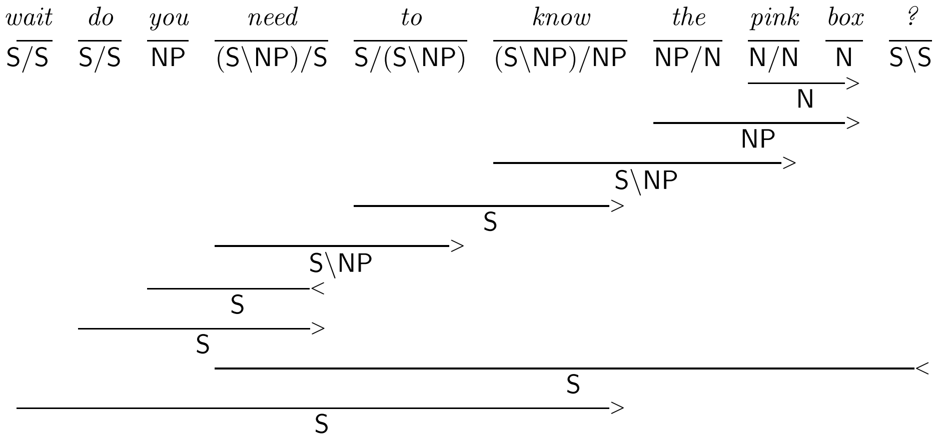
In this section, we present examples for phenomena that distinguish the textual basis of the CReST corpus from the Penn Treebank. Thus, we focus on phenomena typical for spontaneous speech that do not occur in the Penn Treebank or are less frequent there. For the sentences, we present the syntactic analyses in all three syntactic formalisms.

4.1 Questions and Imperatives

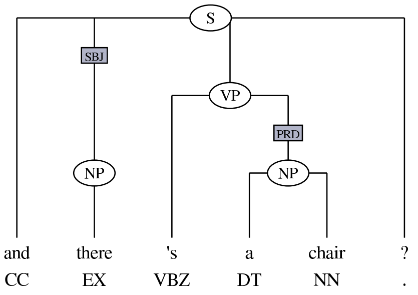
While the Penn Treebank consists mostly of statements, CReST has a significant number of questions and commands: Among the 5 872 sentences, there are 843 questions and 550 commands. In comparison to QuestionBank Judge et al. (2006), CReST has a high number of yes/no questions. The constituent annotation for a typical question, the corresponding dependency and CCG annotation are shown below:



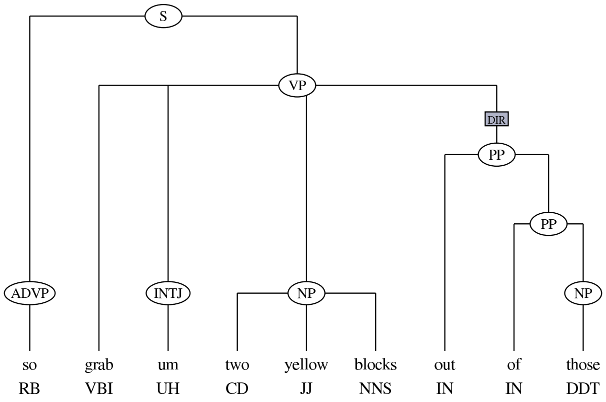


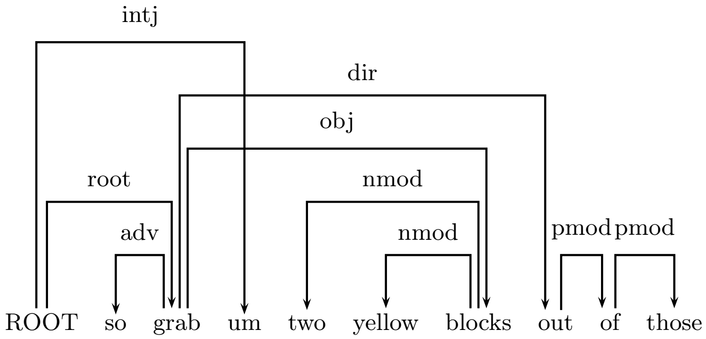


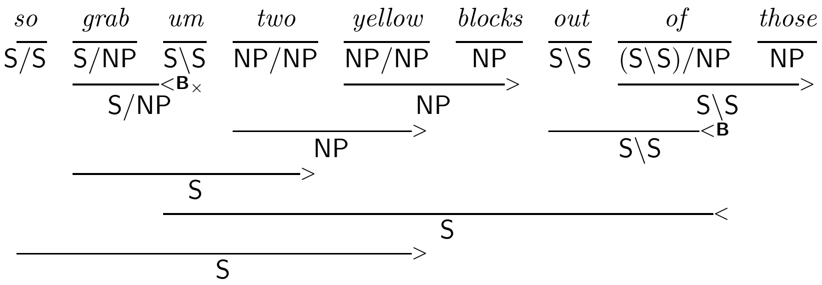
Since CReST is based on cooperative dialogues, many questions are backchannels rather than requests for information. Such questions often have the non-inverted word order of a statement in combination with raising intonation. In the constituent annotation, they are projected to an S node, but they end in a question mark. We show an example below:



An example of a command is shown here:

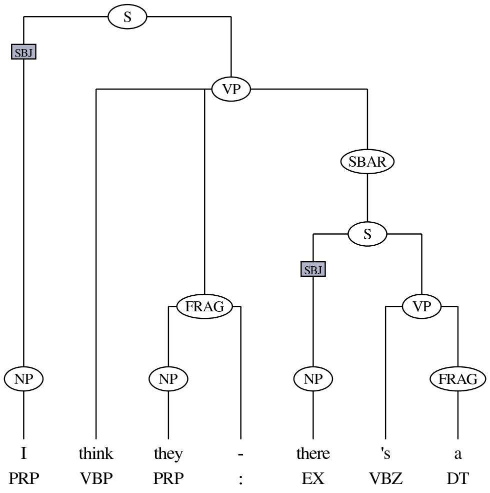


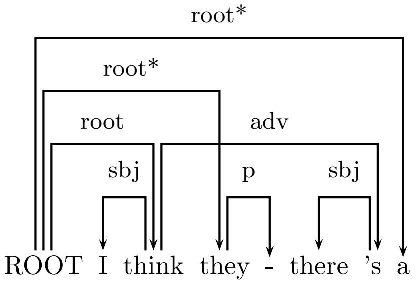


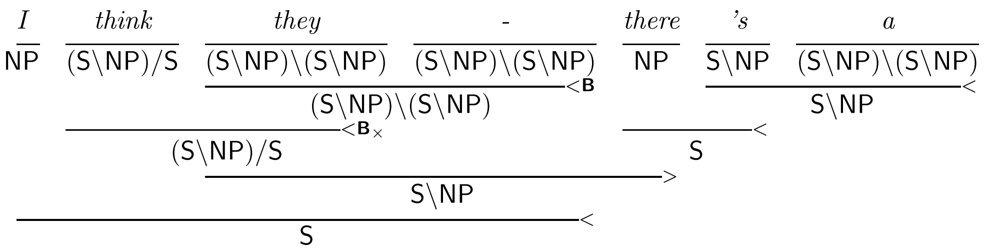

so grab um two yellow blocks out of those
S/S S/NP S\S NP/NP NP/NP NP S\S (S\S)/NP NP
<B×
S/NP
>
NP
>
S\S
>
NP
<B
S\S
>
S
<
S
>
S

4.2 Fragments and Corrections

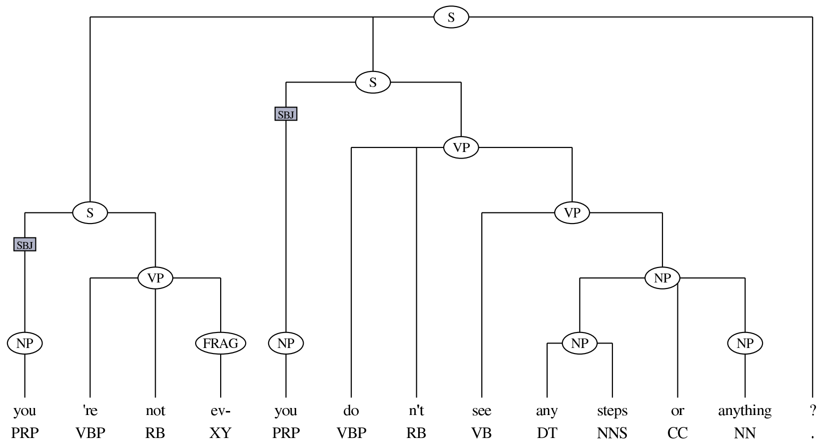
CReST sentences also have a high percentage of fragmented utterances and corrections, which are typical for spontaneous speech. In the constituent annotation, fragments are grouped under a FRAG node and integrated into the remainder of the sentence. The only exception are nonwords, which receive the POS tag XY; those are attached directly to the sentence. In the dependency annotation, fragments remain unattached, and ungrammatical dependencies are starred. Fragments are treated as adjuncts in the CCG annotation, allowing them to join via function combination and then seek a head node. Below, we show a sentence containing two fragments.

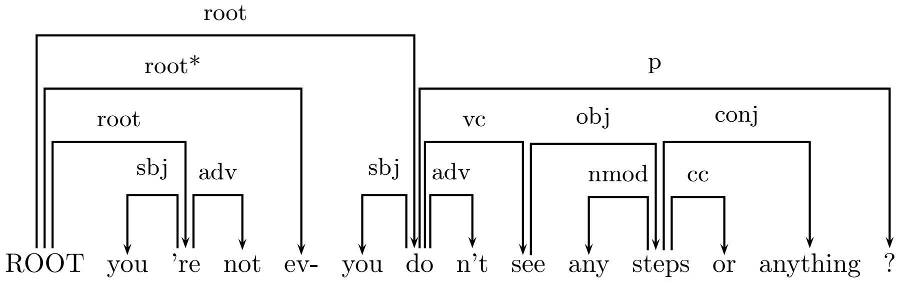






The following shows a sentence containing a correction, the CCG version is shown as the first example in Figure 1.





4.3 Extraposition and Coordination

Spontaneous language often show overt editing or a high compression of information in elliptical constructions. Such phenomena are generally

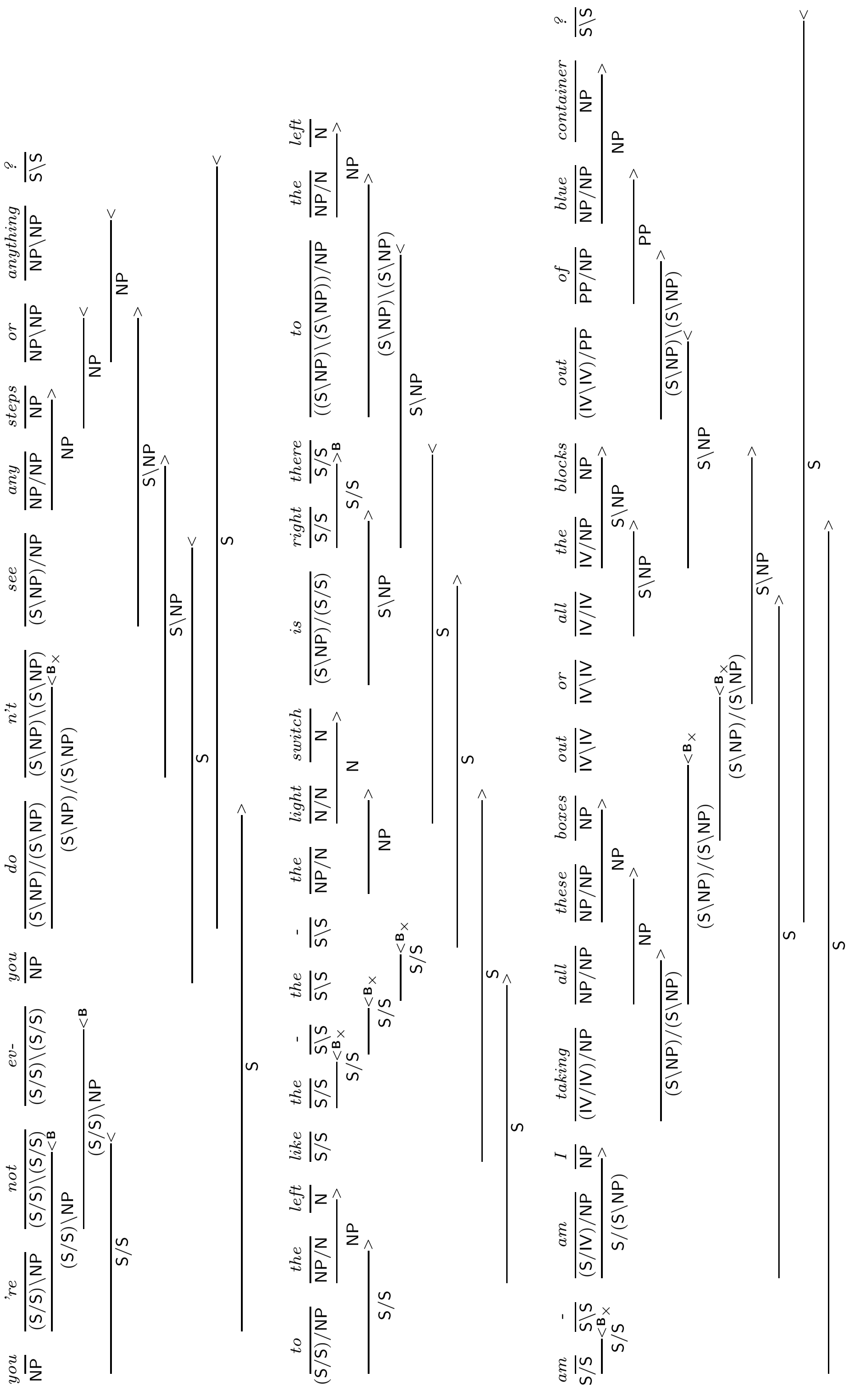
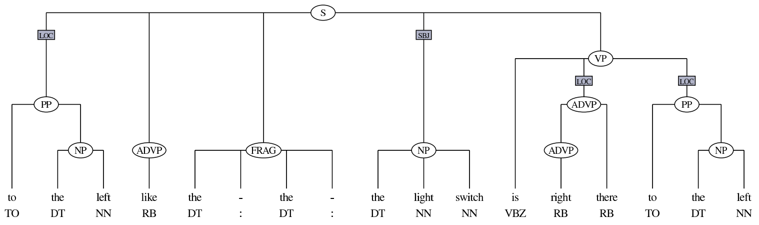
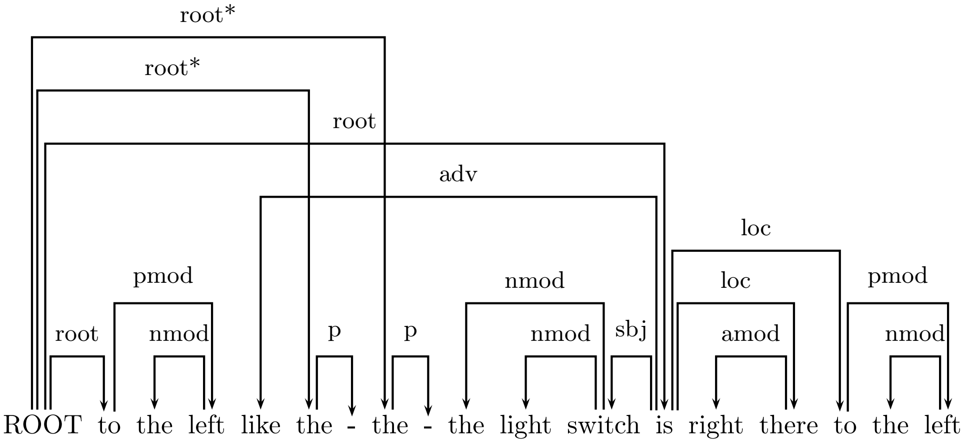


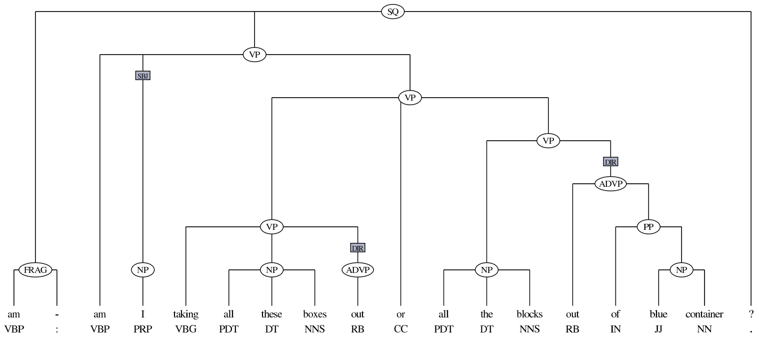
FIGURE 1 The CCG annotations for the examples containing a correction, an extraposition, and an elliptical coordination. In the final example, we abbreviate (S\NP) as IV (intransitive verb).

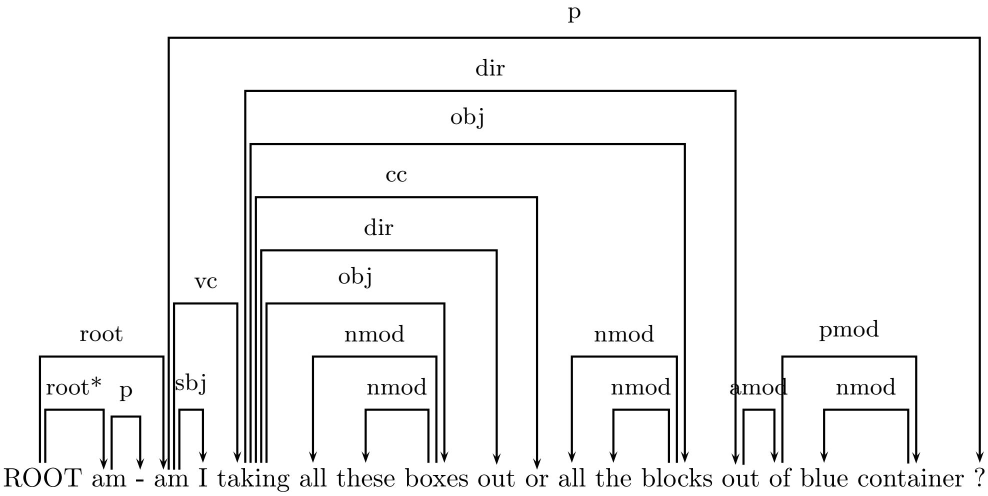
not present in written language, where there is less need for speed and conciseness. Below, we show an example in which a short answer `to the left`, is then elaborated further. In such cases, we decided to treat the first answer as a fronted element of the following clause. The CCG version is shown as the second example in Figure 1.




root*
root*
root
adv
loc
pmod
nmod
loc
pmod
root
nmod
p
p
nmod
sbj
amod
nmod
ROOT to the left like the - the - the light switch is right there to the left

The following illustrates an elliptical coordination. The CCG version is shown as the third example in Figure 1.




p
dir
obj
cc
dir
obj
vc
root
nmod
nmod
pmod
root*
p
sbj
nmod
nmod
amod
nmod
ROOT am - am I taking all these boxes out or all the blocks out of blue container ?

5 Conclusion

We presented the CReST corpus developed from natural language dialogue data collecting as part of a remote search task between two humans as it naturally occurs in a variety of domains. In addition to the audio data, the corpus contains fully transcribed text with disfluency annotations and, for the purpose of this paper most critically, three different syntactic annotations based on constituent, dependency, and combinatory categorial grammar. The corpus is the first of its kind, providing parallel syntactic annotation based on three different grammar formalisms. This parallel annotation allows for the direct comparison and evaluation of linguistic phenomena as well as of parsers based on the three grammar formalisms in an unprecedented way in a naturalistic task. We believe that such comparisons are not only of great utility for the linguistics and computational linguistics community, but also for artificial intelligence and robotics researchers who intend to develop complete natural language understanding systems for agents that are intended to interact with humans in natural ways.

Acknowledgments

This work is based on research supported by the US Office of Naval Research (ONR) Grant #N00014-10-1-0140 and the ONR MURI grant #N00014-07-1-1049.

References

Bos, Johan, Cristina Bosco, and Alessandro Mazzei. 2009. Converting a dependency treebank to a categorial grammar treebank for Italian. In *Proceedings of the Eigth Workshop on Treebanks and Linguistic Theories (TLT-8*. Milan, Italy.

Bosco, Cristina and Vincenzo Lombardo. 2004. Dependency and relational structure in treebank annotation. In *Proceedings of the COLING Workshop on Recent Advances in Dependency Grammar*, pages 9–16.

Carletta, Jean, 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.

Collins, Michael. 1999. *Head-Driven Statistical Models for Natural Language Parsing*. Ph.D. thesis, University of Pennsylvania.

Eberhard, Kathleen, Hannele Nicholson, Sandra Kübler, Susan Gunderson, and Matthias Scheutz. 2010. The Indiana "Cooperative Remote Search Task" (CReST) Corpus. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC)*. Valetta, Malta.

Hockenmaier, Julia and Mark Steedman. 2007. CCGbank: A Corpus of CCG derivations and dependency structures extracted from the Penn Treebank. *Computational Linguistics* 33(3):355–396.

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

Judge, John, Aoife Cahill, and Josef van Genabith. 2006. Questionbank: Creating a corpus of parse-annotated questions. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pages 497–504. Sydney, Australia.

Magerman, David M. 1994. *Natural Language Parsing as Statistical Pattern Recognition*. Ph.D. thesis, Stanford University.

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

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

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