Computational Semantics and Evaluation Benchmark for Interrogative Sentences via Combinatory Categorial Grammar

Hayate Funakura
Kyoto University
Kyoto, Japan
funakura.hayate.28p
@st.kyoto-u.ac.jp

Koji Mineshima
Keio University
Tokyo, Japan
minesima@abelard.flet.keio.ac.jp

Abstract

We present a compositional semantics for various types of polar questions and *wh*-questions within the framework of Combinatory Categorial Grammar (CCG). To assess the explanatory power of our proposed analysis, we introduce a question-answering dataset QSEM specifically designed to evaluate the semantics of interrogative sentences. We implement our analysis using existing CCG parsers and conduct evaluations using the dataset. Through the evaluation, we have obtained annotated data with CCG trees and semantic representations for about half of the samples included in QSEM. Furthermore, we discuss the discrepancy between the theoretical capacity of CCG and the capabilities of existing CCG parsers.

1 Introduction

Interrogative sentences, encompassing various question types, hold a crucial position in the study of syntax and semantics within the field of theoretical linguistics (Dayal, 2016). Of particular significance are *wh*-questions, which serve as a benchmark for testing linguistic theories that explore the interface between syntax and semantics, including Categorial Grammar (Steedman, 1996). For example, the extraction phenomena involved in *wh*-questions, one of the representative examples in the mismatch between syntax and semantics, provide valuable insights for understanding this interface (Kubota and Levine, 2020). However, despite their importance, the exploration of interrogative sentences within the framework of Categorial Grammar remains relatively underdeveloped with few exceptions (Vermaat, 2005; Xiang, 2021).

Furthermore, while computational linguistics has witnessed growing research on question sentences in terms of semantic parsing (Kwiatkowski et al., 2011; Reddy et al., 2014), there exists a notable disparity between the semantic parsing literature and theoretical investigations into the syntax-semantics interface. The latter research focuses on formal semantics and its detailed examination of various semantic phenomena. This disparity presents an opportunity for bridging the gap and fostering a more integrated approach to the study of interrogative sentences.

Motivated by these gaps in the current literature, this paper aims to present a compositional analysis of different types of interrogatives, including polar and *wh*-questions, within the framework of Combinatory Categorial Grammar (CCG) (Steedman, 2000). This analysis defines a procedure to assign logic-based semantic representations to both questions and their answers, based on their respective CCG trees. These representations can be combined with automated theorem provers to perform logical inferences for question-answering.[1]

To facilitate practical implementation and empirical testing, a computational system ccg2hol will be introduced in this paper. This system leverages existing CCG parsers and can be employed for question-answering tasks by integrating it with a theorem prover.

In order to evaluate the syntactic and semantic analyses of interrogative sentences, we design and introduce a dataset of Question-Answer pairs, which we call QSEM.[2] The construction of this dataset follows the methodological approach established by FraCaS (Cooper et al., 1996), which serves as a reliable starting point for natural language inferences that carefully separates the semantic and pragmatic factors involved in determining entailment relations. The QSEM dataset comprises two primary categories of problems: complex and diverse issues frequently discussed in formal semantics, such as generalized quantifiers and

[1] As will be mentioned later, we reduce question-answering to recognizing textual entailment. Therefore, a theorem prover can be used as a question-answering engine.

[2] QSEM is available at https://github.com/hfunakura/qsem.

scope ambiguity, and problems that are closer to real-world language use commonly observed in question-answering contexts. The former was created based on the FraCaS problems, while the latter was developed using SQuAD v2.0 (Rajpurkar et al., 2018) training data as a basis. The dataset will provide a valuable resource for detailed examination and analysis of the semantic entailment output by the implemented system.

By undertaking this investigation, we aim to not only contribute to the understanding of interrogative sentences within the context of Categorial Grammar but also shed light on the challenges and limitations of the existing CCG parser based on CCGBank (Hockenmaier and Steedman, 2007). Through a thorough examination of interrogative sentences, this study tries to enhance our comprehension of the syntax, semantics, and computational aspects involved, thereby offering valuable insights for future research and applications in the field of computational linguistics and related fields.

What we prioritize most is the establishment of the system `ccg2hol`, which makes theoretical linguistics computationally implementable. In addition to that, our contributions lie in the following four aspects.

1. We present a compositional analysis that maps various types of interrogative sentences to logical semantic representations within the framework of CCG.
2. We introduce a FraCaS-inspired benchmark `QSEM` for evaluating the syntax-semantics interface for various types of interrogative sentences.
3. We report a semantic annotation project which assigns each sentence in `QSEM` with a gold CCG tree and a logical semantic representation using our system `ccg2hol`.
4. We perform a qualitative analysis of the output from standard CCG parsers.

The paper is structured as follows. In Section 2 we introduce some background in formal semantics and computational semantics of interrogative sentences. In Section 3, we present our analysis of the syntax and semantics of interrogative sentences in CCG. In Section 4, we provide an overview of the `QSEM` dataset and its characteristics. In Section 5, we introduce our semantic composition and logical inference system `ccg2hol`, which is based on existing CCG parsers and automated theorem provers. We also describe the evaluation and annotation using this system, as well as the qualitative analysis of the CCG parsers.

2 Related work

The compositional semantics of interrogative sentences began with Hamblin (1973) and Karttunen (1977), and has been developed by subsequent researches (Groenendijk and Stokhof, 1984; Krifka, 2001; Ciardelli et al., 2019, etc.). Research in this area covers a wide range of topics, including question-answer relationships, presuppositions, and scope problems. Also, various phenomena related to the embedding of interrogative clauses are being actively addressed. While various proposals have been made for specific phenomena and constructions, it is not clear how to test the applicability of each analysis to a wide range of interrogative sentences. It is an important question whether the analyses proposed within a given paper are valid only in a very limited number of cases or whether they have a high degree of generality.

This situation is not limited to questions, but to formal semantics in general. FraCaS (Cooper et al., 1996) is an early benchmark proposed as a basis for systematically evaluating proposals in formal semantics and there have been several subsequent test sets proposed for the evaluation of formal semantics since then, including MultiFraCaS project[3] and JSeM (Kawazoe et al., 2017).[4] Watanabe et al. (2019) provide a dataset for evaluating the semantics of questions, including examples of *wh*-questions, polar questions, and alternative questions. However, the dataset has limitations in variation, as it does not include *wh*-words other than *who*, and there are no instances where the object is a *wh*-word. To the best of our knowledge, there is no inference test suite that covers a broader range of linguistic constructions and phenomena related to questions than `QSEM`.

3 Syntax and semantics

We give our analysis to the following types of questions:

- Polar questions
- Argument *wh*-questions (*who*, *what*, *which*)
- Adjunct *wh*-questions (*when*, *where*)

[3] https://gu-clasp.github.io/multifracas/
[4] https://github.com/DaisukeBekki/JSeM

We seek here to account for the question-answer relationship. Other semantic phenomena associated with questions include presuppositions, ambiguity in question-embedded sentences, and the anaphoric nature of polarity particles. We limit our account here to the following question-response pairs, where the goal is to describe that the response is the answer to the question.

(1) Polar questions
  a. Did John meet Mary?
  b. John met Mary.

(2) Argument *wh*-questions
  a. Who smokes?
  b. John smokes.

(3) *When*-questions
  a. When did John meet Mary?
  b. John met Mary yesterday.

(4) *Where*-questions
  a. Where did John meet Mary?
  b. John met Mary at the station.

We define the relationship between questions and answers in terms of entailment and contradiction relations. In other words, our theory predicts that a response is an answer to a question when the semantic representation of the response entails or contradicts the semantic representation of the question.

The language for semantic representation is a higher-order logic language (Mineshima et al., 2015), combined with event, time, and location variables. Intuitionistic logic is assumed as the logical system, and Coq is used as the inference engine accordingly. Section 3.1 provides examples of the semantic representations assigned to each type of interrogative. Section 3.2 discusses the derivation of semantic representations by CCG.

3.1 Semantic representations for questions

The following are examples of the semantic representations we assign to each type of interrogative sentence.

(5) Polar questions
  a. Did John meet Mary?
  b. $?(\exists x.\exists y.\exists e.[\mathrm{John}(x) \wedge \mathrm{Mary}(y) \wedge \mathrm{Meet}(e) \wedge \mathrm{Subj}(e,x) \wedge \mathrm{Obj}(e,y)])$

(6) Argument *wh*-questions
  a. Who smokes?
  b. $\mathsf{Q}(\lambda x.[\mathrm{Smoke}(x)])$

(7) *When*-questions
  a. When did John meet Mary?
  b. $\mathsf{Q}(\lambda t.\exists x.\exists y.\exists e.[\mathrm{John}(x) \wedge \mathrm{Mary}(y) \wedge \mathrm{Meet}(e) \wedge \mathrm{Subj}(e,x) \wedge \mathrm{Obj}(e,y) \wedge \mathrm{TimeOf}(e,t)])$

(8) *Where*-questions
  a. Where did John meet Mary?
  b. $\mathsf{Q}(\lambda l.\exists x.\exists y.\exists e.[\mathrm{John}(x) \wedge \mathrm{Mary}(y) \wedge \mathrm{Meet}(e) \wedge \mathrm{Subj}(e,x) \wedge \mathrm{Obj}(e,y) \wedge \mathrm{LocOf}(e,l)])$

What exactly the operators ? and Q should be is a purely semantic question.

No matter how they are defined, there is no effect on semantic composition. Since our goal is to establish a semantic composition workflow consistent with the CCG parsers, we define ? and Q in a very simple form.

(9) a. $?(P) \equiv P \vee \neg P$
  (where P is a formula of type t)
  b. $\mathsf{Q}(f) \equiv \exists x.f(x)$
  (where f is a first-order function)

The above representations are partially based on those of inquisitive semantics (Ciardelli et al., 2019); for polar questions they are the same as in inquisitive semantics, while for wh-questions, we discard the ambiguity about exhaustivity that is considered in Ciardelli et al. (2019), thus simplifying the treatment of inquisitive semantics.

There are various alternative options for defining ? and Q. We mention two of them. First, a Karttunen-style analysis can be achieved by defining these operators as follows:

(10) $?(P) \equiv \lambda p.[p(w_a) \wedge p = \lambda w.p(w)]$

(11) $\mathsf{Q}(f) \equiv \lambda p.\exists x.[p(w_a) \wedge p = \lambda w.f(x)(w)]$

Here, w_a denotes a designated (actual) world.

Second, it is also possible to define the ? and $\mathcal{Q}$ operators in terms of modal logic, which enables to express three readings with respect to exhaustivity by providing three $\mathcal{Q}$ operators (Nelken and Shan, 2004, 2006).

(12) Semantic representation of questions using modality
  a. $\mathsf{Q}_{ms}(f) = \exists x.[\Box f(x)]$
  (mention-some reading)

b. $\mathsf{Q}_{we}(f) = \forall x.[f(x) \rightarrow \Box f(x)]$
(weakly exhaustive reading)

c. $\mathsf{Q}_{se}(f) = \forall x.[(f(x) \rightarrow \Box f(x)) \wedge (\neg f(x) \rightarrow \Box \neg f(x))]$
(strongly exhaustive reading)

As discussed above, depending on how ? or Q are defined, this analysis can embody various perspectives. We do not intend to commit to a specific position. Therefore, we adopt (9-a) and (9-b) for simplicity. Note that we have chosen intuitionistic logic as the underlying logic, mainly because of its compatibility with theorem provers.

3.2 Compositional semantics

In this subsection, we present the lexical items defined for the words that play a central role in our analysis: *be*, *do*, and *wh*-words. We also demonstrate semantic composition using them.

3.2.1 Lexical entries

The lexical entries for *be* and *do* are shown in Table 1.[5] We assume that *be* and *do* appearing in interrogative sentences are distinct in the lexicon from those appearing in declarative sentences. In other words, separate lexical entries are defined for *be* and *do* that appear in declarative sentences (omitted here).

The lexical entries for *wh*-words are shown in Table 2. The morphemes when_2 and where_2 are assumed to appear in the construction `wh-be-NP` (see examples below).

(13) When is the deadline?

(14) Where is the office?

3.2.2 Semantic composition

In this subsection, we provide examples of semantic composition for both *wh*-questions and polar questions.

Figure 1 shows an example of semantic composition for a *wh*-question. Since we assume that the verb introduces event quantification, we use the Quantifier Closure rule (**QC**) in addition to the CCG combinatory rules. **QC** is a unary rule, which applies the input expression to $\lambda x.\top$.

An example of semantic composition for polar questions is shown in Figure 2. To derive the semantic representation of polar questions, we define a CCG unary rule, ?**I**, which is a rule that transforms an expression P into $?P$, through which the semantic representation of a polar question is obtained.

[5]We have not attributed the ? operator appearing in the semantic representation of polar questions to the lexical meaning of *be* or *do*, but have defined a unary rule of CCG that introduces the ? operator. This is more of a practical measure for ease of implementation rather than a theoretical one.

4 Dataset

`QSEM` consists of questions and the responses to those questions. The primary goal in creating this dataset was to establish a basis for quantitatively measuring the degree of agreement between the predictions derived from semantic representations and our intuitive judgments. What this dataset asks the system is whether a given answer qualifies as an answer to the question.

The following are examples of the samples included in the dataset. "P" represents the premise and "Q" represents the question. Labels have three possible values: `yes`, `no`, and `unknown` (the rules for label assignment are discussed in Section 4.3).

(15) ID: 6
P1 Every Italian man wants to be a great tenor.
Q Who wants to be a great tenor?
Label: yes

(16) ID: 35
P1 No delegate finished the report on time.
Q Which delegate finished the report on time?
Label: no

(17) ID: 72
P1 Amish separated from the Mennonites in 1693.
Q When did the Anabaptists split?
Label: unknown

The format of our dataset is based on FraCaS (Cooper et al., 1996), a pioneering semantic evaluation dataset. Before going into the details of our dataset, the following subsection provides an overview of FraCaS as a background.

4.1 Background: FraCaS

FraCaS is a test suite for evaluating NLP systems and linguistic theories. The first version provided in Cooper et al. (1996) contains one or more assumptions, a polar question, and an answer to that question (*yes*, *no*, *don't know*, etc.). There are

Expression	Category	Semantics
be_1	$(S_q/NP)/NP$	$\lambda P_1 P_2 K.P_2(\lambda y.\top, \lambda x.Q_1(\lambda y.\top, \lambda y.\exists e.[\text{Be}(e) \wedge (\text{Subj}(e) = y) \wedge K(e)])$
be_2	$(S_q/(S_{adj}\backslash NP))/NP$	$\lambda P_1 P_2 K.P_2(\lambda y.\top, \lambda x.P_1(\lambda y.\top, \lambda y.\exists e.[\text{Be}(e) \wedge (\text{Subj}(e) = y) \wedge K(e)]))$
be_3	$(S_q/(S_{pss}\backslash NP))/NP$	$\lambda P_1 P_2 K.P_2(P_1, \lambda e.K(e))$
do	$(S_q/(S_b\backslash NP))/NP$	$\lambda P_1 P_2 K.P_2(P_1, K)$

Table 1: Lexical entries for *be* and *do*

Expression	Category	Semantics
who	$S_{wq}/(S\|NP)$	$\lambda PK.\mathsf{Q}(\lambda x.P(\lambda F_1 F_2.F_2, \lambda y.\top))$
what_1	$S_{wq}/(S\|NP)$	$\lambda PK.\mathsf{Q}(\lambda x.P(\lambda F_1 F_2.F_2, \lambda y.\top))$
what_2	$(S_{wq}/(S\|NP))/N$	$\lambda P_1 P_2 K.\mathsf{Q}(\lambda x.[P_1(x) \wedge P_2(\lambda F_1 F_2.F(x), \lambda y.\top)])$
which	$(S_{wq}/(S\|NP))/N$	$\lambda P_1 P_2 K.\mathsf{Q}(\lambda x.[P_1(x) \wedge P_2(\lambda F_1 F_2.F(x), \lambda y.\top)])$
when_1	S_{wq}/S_q	$\lambda SK.\exists t.\mathsf{Q}(S(\lambda e.\text{TimeOf(e, t)}))$
when_2	$S_{wq}/(S_q/NP)$	$\lambda PK.\mathsf{Q}(\lambda t.P(\lambda F_1 F_2.(F_1 \wedge F_2), \lambda e.\exists t.\text{TimeOf}(e, t)))$
where_1	S_{wq}/S_q	$\lambda SK.\exists l.\mathsf{Q}(S(\lambda e.\text{LocOf}(e, l)))$
where_2	$S_{wq}/(S_q/NP)$	$\lambda PK.\mathsf{Q}(P(\lambda F_1 F_2.(F_1 \wedge F_2), \lambda e.\exists l.\text{LocOf}(e, l)))$

Table 2: Lexical entries for *wh*-expressions. Here, we bundle $S_{dcl}\backslash NP$ and S_q/NP together and denote them as $S|NP$.

346 problems, divided into sections for each phenomenon. And it is controlled not to include difficulties other than the phenomenon in focus. This makes it easy to estimate the explanatory power of the analysis by phenomenon.

The following is an example of an original problem:

(18) ID: 3.1 in Cooper et al. (1996)
P1 An Italian became the world's greatest tenor.
Q Was there an Italian who became the world's greatest tenor?
Label: yes

The original form was thus a dataset consisting of QA pairs, but then hypotheses (H) were added by Bill MacCartney and formulated as a set of implication recognizing textual entailment.[6] The following are examples of questions from the latest version of FraCaS:

(19) ID: fracas-001
P1 An Italian became the world's greatest tenor.
Q Was there an Italian who became the world's greatest tenor?
H There was an Italian who became the world's greatest tenor.
Label: yes

(20) ID: fracas-085
P1 Exactly two lawyers and three accountants signed the contract.
Q Did six lawyers sign the contract?
H Six lawyers signed the contract.
Label: no

(21) ID: fracas-117
P1 Every student used her workstation.
P2 Mary is a student.
Q Did Mary use her workstation?
H Mary used her workstation.
Label: yes

[6]This version is available at https://nlp.stanford.edu/~wcmac/downloads/fracas.xml

Type of Question	**Count**
Polar	23
Who	15
What	20
Which	22
When	35
Where	23

Table 3: The number of samples for each type of question

FraCaS includes a wide range of topics, including tense, anaphora, and propositional attitudes. In addition, all samples include polar questions. However, there is no section focusing on the semantic behavior of the questions themselves. Also, to our knowledge, there are few other evaluation datasets for question semantics. This motivates our proposed dataset. In the following subsections, we describe the contents of our dataset and the process of its creation.

4.2 Dataset organization

Our dataset consists of 138 samples. We place more emphasis on the qualitative aspects, such as

$$\dfrac{\dfrac{\text{Who}}{\begin{array}{c}S_{wq}/(S_{dcl}\backslash NP)\\ \lambda PK.\mathsf{Q}(\lambda x.P(\lambda F_1F_2.F_2,\lambda y.\top))\end{array}} \quad \dfrac{\text{smokes}}{\begin{array}{c}S_{dcl}\backslash NP\\ \lambda PK.P(\lambda x.\top,\lambda x.\exists e.[\text{Smoke}(e)\wedge\text{Subj}(e,x)\wedge K(e)])\end{array}}\ >}{\begin{array}{c}S_{wq}:\lambda K.\mathsf{Q}(\lambda x.\exists e.[\text{Smoke}(e)\wedge\text{Subj}(e,x)])\\ \bar{S}_{wq}:\mathsf{Q}(\lambda x.\exists e.[\text{Smoke}(e)\wedge\text{Subj}(e,x)])\end{array}}\ \mathbf{QC}$$

Figure 1: An example of semantic composition for a *wh*-question. To ensure Categorial Type Transparency (Steedman, 2000), the category derived by **QC** is distinguished from S_{wq} and is denoted as $\bar{S}_{wq}$.

$$\dfrac{\dfrac{\dfrac{\text{Does}}{\begin{array}{c}(S_q/(S_b\backslash NP))/NP\\ \lambda P_1P_2K.Q_2(Q_1,K)\end{array}} \quad \dfrac{\dfrac{\dfrac{\text{John}}{N}}{\begin{array}{c}\text{John}\\ NP\end{array}}\ \mathbf{TR}}{\lambda F_1F_2.F_1(\text{John})\wedge F_2(\text{John})}}{\begin{array}{c}S_q/(S_b\backslash NP)\\ \lambda P_2K.Q_2(\lambda F_1F_2.F_1(\text{John})\wedge F_2(\text{John}),K)\end{array}}\ > \quad \dfrac{\dfrac{\dfrac{\text{like}}{\begin{array}{c}(S_b\backslash NP)/NP\\ \cdots\end{array}} \quad \dfrac{\dfrac{\dfrac{\text{Smith}}{N}}{\begin{array}{c}\text{Smith}\\ NP\end{array}}\ \mathbf{TR}}{\lambda F_1F_2.F_1(\text{Smith})\wedge F_2(\text{Smith})}}{S_b\backslash NP}\ >}{\begin{array}{c}S_q\\ \lambda P_2K.P_2(\lambda y.\top,\lambda x.\exists e.[\text{Like}(e)\wedge(\text{Subj}(e)=x)\wedge(\text{Obj}(e)=\text{Smith})\wedge K(e)])\end{array}}\ >$$

$$\dfrac{\begin{array}{c}\bar{S}_q\\ \lambda K.\exists e[\text{Like}(e)\wedge(\text{Subj}(e)=\text{John})\wedge(\text{Obj}(e)=\text{Smith})\wedge K(e)]\end{array}}{\begin{array}{c}\bar{S}_{pol}\\ \exists e[\text{Like}(e)\wedge(\text{Subj}(e)=\text{John})\wedge(\text{Obj}(e)=\text{Smith})]\end{array}}\ \mathbf{QC}\ \ ?\mathbf{I}$$

$$?\exists e[\text{Like}(e)\wedge(\text{Subj}(e)=\text{John})\wedge(\text{Obj}(e)=\text{Smith})]$$

Figure 2: An example of semantic composition for a polar question

the variety of constructions and phenomena, and the accuracy of labels, rather than quantitative aspects.

Table 3 shows the number of samples for each type of question. Each sample is annotated as to which type of question is it related to. Thus, it is easy to measure the degree to which the system is applicable to which type of questions. A wider range of questions, such as alternative questions, *how*-questions, *why*-questions, etc., will need to be covered in the future.

The dataset consists of the following problems:

1. Problems that test understanding of quantificational expressions
2. Problems that test syntactic and semantic understanding of multiple *wh*-questions
3. Problems that test understanding of the interaction of quantifiers and *wh*-word scopes
4. Problems that test comprehension of basic *wh*-questions

1-3 focuses on whether the system can solve semantically challenging problems. 4, on the other hand, focuses on the general applicability of proposed analyses. Each of the above four types of questions was created from different resources. In the following subsection, we will explain the process of creating each question, presenting sample examples.

4.3 Dataset creation process

In this subsection, we describe the labeling rules and how we collected the QA pairs.

Labeling rules Each QA pair in `QSEM` is assigned one of the labels: `yes`, `no`, or `unknown`. The rules for label assignment are as follows.

- Problems copied from FraCaS
  - We use the original labels assigned in FraCaS.
- Problems created by the authors
  - When the premises directly answer the question, `yes` is assigned.
  - When the premises negate the presupposition of the question, `no` is assigned.
  - When none of the above conditions apply, `unknown` is assigned.

Quantificational expressions Pairs of polar questions and responses were extracted from sections 1.1 and 1.2 of FraCaS as samples for quantification. The current version includes only those examples in which either *every*, *all*, *each*, *some*, *a*, or *no* is used. Of the 23 polar questions, 14 were copied from FraCaS. The other 9 were created by the authors based on the FraCaS samples.

Multiple *wh*-questions The interrogative sentences in which both a fronted *wh*-phrase and a *wh*-phrase in-situ occur together, i.e., multiple *wh*-questions, is also included in this dataset. This construction is ambiguous between single-pair reading

and pair-list readings. There are vigorous debates to explain this ambiguity.

The QA pairs on multiple *wh*-questions are taken from Dayal (2016). The collected samples are shown below.

(22) ID: 40
P1 Bill met Carl.
P2 Bill is a student.
P3 Carl is a professor.
Q Which student met which professor?
Label: yes

(23) ID: 41
P1 Bill met Carl and Alice met Dan.
P2-3 Bill is a student., Alice is a student.
P4-5 Carl is a professor., Dan is a professor.
Q Which student met which professor?
Label: yes

P1 and Q are copied from the source literature, while the premises after P2 are added by the authors to provide a context.

Scope ambiguity *wh*-interrogatives in which quantificational expressions occur are also a central subject of study in this area.

(24) Who does everyone like?
a. Tell me about one person who is liked by all. (wh>∀)
b. For each person, tell me who that person likes. (∀>wh)

On the other hand, such ambiguity does not arise when the quantificational expression appears in the object position.

(25) Who likes everyone? (wh>∀)

Samples for scope ambiguity were taken from Chierchia (1993) and Krifka (2003). Examples are shown below.

(26) ID: 48
P1 Bill likes Smith and Sue likes Jones.
Q Who does everyone like?
Label: yes

(27) ID: 49
P1 Everyone likes Smith.
Q Who does everyone like?
Label: yes

Type of Question	Count
Who	10
What	10
Which	8
When	35
Where	23

Table 4: The number of samples obtained from SQuAD by question type

(28) ID: 44
P1 Bill likes Smith and Alice likes Jones.
P2-3 Bill is a student., Alice is a student.
P4-5 Smith is a professor., Jones is a professor.
Q Which student likes every professor?
Label: no

Basic *wh*-questions To test for a greater variety of constructions, samples were created based on SQuAD training data. SQuAD is a question-answering dataset that is often used as a benchmark for natural language processing systems. The dataset contains approximately 90K QA pairs, and the system must provide an answer to a question based on the content of a given paragraph. In addition to the above samples, there are about 40K questions for which no answer can be found from the given paragraphs.

We performed random sampling by question type from the questions in this dataset. From them, we excluded samples in which the following phenomena and constructions were critically involved.

- idiom
- coordination
- anaphora
- tense
- degree

We also excluded questions on sensitive topics. We performed the above work with a random sampling size of 50 for the *when* and *where* questions, and 30 for the *who*, *what*, and *which* questions. The answers included in SQuAD are so-called non-sentential answers. Based on these, we created answers for QSEM. Table 4 shows the final sample size obtained for each type of interrogative.

Limitations Here, we will discuss the main limitations of QSEM.

QSEM includes examples related to negative interrogative sentences (as shown below), but it does not capture the fact that such questions often receive rhetorical interpretations.

(29) ID: 115
P1 Constantius did not consent to a new trial.
Q Who did not consent to a new trial?
Label: yes

Additionally, while focus plays a crucial role in the relationship between a question and its answer, the current QSEM abstracts away from information related to focus.

5 Implementation and semantic annotation

This section provides an overview of the annotation and the results of the evaluation.

5.1 Pipeline

In this subsection, we describe our system ccg2hol for semantic composition and logical inference. The pipeline for this system is shown in Figure 4. Among the components in the diagram, what we implemented is the semantic composition system that takes CCG derivation trees as input and outputs formulas (HOLs), and the interface between each component.

The system first takes one or more sentences as input and performs syntactic parsing using existing CCG parsers. C&C parser (Clark and Curran, 2007), EasyCCG (Lewis and Steedman, 2014), and depccg (Yoshikawa et al., 2017) are used as CCG parsers. Based on the results of the parsing, a semantic tag is assigned to each word. Then, the semantic composition is performed using the CCG tree and semantic tags (Abzianidze and Bos, 2017).[7] As a semantic representation, we propose an abstract expression that is independent of specific semantic analysis. We call this expression HOL (higher-order logic). HOL has information on syntactic dependencies and semantic tags (e.g., Figure 3). At present, we are mechanically assigning semantic tags from CCG categories, POS tags, NER tags, and lemmas. Semantic tags were proposed as a superior resource for judging lexical meaning than POS tags or NER tags, so assigning semantic tags based on these pieces of information is not ideal. Therefore, it is desirable for semantic tags to be determined by an independent assigner. In this study, as a provisional measure, we manually supplemented areas where POS tags or NER tags were insufficient (see Section 5.2 below).

HOL can be converted into specific expressions such as FOL, DRT, etc. Therefore, by utilizing HOL as the primary semantic representation, our system can be used independently of any specific theoretical framework or analytical strategy. An example of HOL is shown in Figure 3. HOLs are then converted into FOL expressions for inference. The FOL expressions are passed to the theorem prover Coq (Bertot and Castéran, 2004) to perform inference and predict entailment and contradiction relations.

5.2 Evaluation and Annotation

Using the pipeline described above, we evaluated the degree to which our analysis could address the QSEM problem. And we have accumulated samples that contain no errors in HOLs or inference results as gold data. The main source of errors is in deriving HOLs. These errors mainly fall into two categories: mistakes in CCG trees and inaccuracies in semantic tag assignment. If an error could be resolved by simply adjusting the semantic tags, we manually made the corrections.

In the following, we will discuss the manually assigned semantic tags and provide a qualitative error analysis of the CCG parsers. Lastly, we will report on the extent of completed annotations within the QSEM data.

Semantic tags To represent the polysemy of prepositions, we manually corrected the output of the system. Using only the information utilized in the above-mentioned pipeline for semantic tagging, we could not differentiate between prepositions used for time and those used for location, leading to the same semantic tag being assigned in all cases. This resulted in difficulties when dealing with examples of *when* and *where* questions. Therefore, we manually assigned different semantic tags to time prepositions appearing in expressions like *on December 12* and location prepositions appearing in expressions like *in Oxford*.

Main errors in CCG trees Through the observation of our analysis results, the following tendencies in existing CCG parsers were suggested:

[7] While we have followed the idea of (Abzianidze and Bos, 2017) to use semantic tags as a key to determine lexical meanings, the specific design of the tag set was carried out by ourselves.

$$((\text{which}_{\mathsf{WDT}}\ \text{delegate}_{\mathsf{NN}})\ (\text{finish}_{\mathsf{TV}}\ (\text{the}_{\mathsf{DT}}\ \text{report}_{\mathsf{NN}})))$$

Figure 3: HOL corresponding to *Which delegate finished the report?*

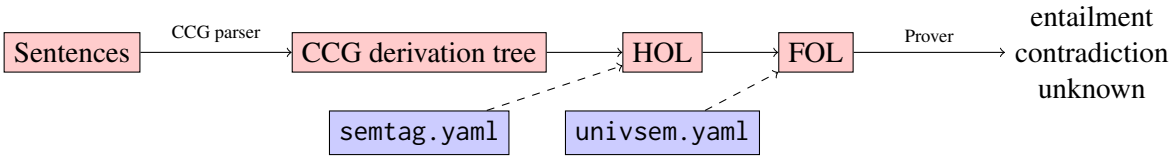

Figure 4: Pipeline of `ccg2hol`

- Multiple *wh*-questions are particularly difficult for parsers.
- There is a tendency for the analysis of past participles to be inconsistent between declarative and interrogative sentences.

The current version of `QSEM` includes one instance of a multiple-*wh* question.

(30) ID: 40, 41
Which student met which professor?

All parsers we employed failed in analyzing this instance. C&C parser and depccg identified *which professor* as an embedded interrogative clause. EasyCCG analyzed the two *which* as if they were adjectives, and recognized the entire sentence as a declarative sentence.

In addition, discrepancies were observed between declarative and interrogative sentences regarding past participles appearing as complements to *be*, such as *located* in the following example.

(31) ID: 129
Q Where is Symphony Hall located?
A Symphony Hall is located on the west of Back Bay.

There are 16 instances in `QSEM` that involve the use of $\mathsf{be\text{-}V_{pp}}$. For 12 of these, both C&C and depccg recognized the past participle form appearing in interrogatives as an adjective, while recognizing the past participle form appearing in declaratives as the passive voice of a verb. Even in cases where the analysis of the past participle was consistent, there were errors in other parts of the tree. EasyCCG was consistent, recognizing both types of past participle as likely being in the passive voice. However, there was not a single instance where the entire tree was correctly parsed.

Results We performed `ccg2hol` analysis for each sample in `QSEM`, and considered those that correctly produced CCG trees, semantic representations, and inference results as gold data for annotation. Currently, annotations have been completed for approximately 49.3% (68 out of 138) of the entire `QSEM`. Many of the analyses yet to be annotated include results with parsing errors from the CCG parsers as mentioned above, and results with inaccurate semantic tags assigned.

6 Conclusion

In this paper, we proposed an extensive analysis of the interrogative sentences and proposed a benchmark `QSEM` to evaluate it. In addition, we introduced a system, `ccg2hol`, to implement the proposed analysis. This system was used to annotate a portion of the examples in `QSEM` with CCG trees and HOL.

`QSEM` aims to formulate interesting problems in question semantics as question-answering and will be further augmented in the future. `ccg2hol` is a semantic composition and inference system. The HOL obtained as a result of semantic composition is an abstract structure independent of any specific analysis and, together with the annotated data, can be used for testing various syntactic and/or semantic frameworks.

Our ultimate goal is for `ccg2hol` to be a language-universal and analysis-independent computational framework. To make `ccg2hol` a universal inference and evaluation framework, broader annotations and extensions to other languages are necessary. For wider annotation and an improved inference system, the immediate future challenges to tackle are the elimination of parsing errors by the CCG parsers and the refinement of semantic tag design. For `ccg2hol` to handle other languages,, it is necessary to connect the semantic composition system to parsers for languages other than English.

Acknowledgment

This work is partially supported by JST, CREST grant number JPMJCR2114.

References

Lasha Abzianidze and Johan Bos. 2017. Towards universal semantic tagging. In *International Conference on Computational Semantics*.

Yves Bertot and Pierre Castéran. 2004. *Interactive Theorem Proving and Program Development: Coq'Art: The Calculus of Inductive Constructions*. Springer Science & Business Media.

Gennaro Chierchia. 1993. Questions with quantifiers. *Natural Language Semantics*, 1(2):181–234.

Ivano Ciardelli, Jeroen Groenendijk, and Floris Roelofsen. 2019. *Inquisitive Semantics*. Oxford University Press.

Stephen Clark and James R Curran. 2007. Wide-coverage efficient statistical parsing with CCG and log-linear models. *Computational Linguistics*, 33(4):493–552.

Robin Cooper, Dick Crouch, Jan Van Eijck, Chris Fox, Johan Van Genabith, Jan Jaspars, Hans Kamp, David Milward, Manfred Pinkal, Massimo Poesio, et al. 1996. Using the framework. Technical report, Technical Report LRE 62-051 D-16, The FraCaS Consortium.

Veneeta Dayal. 2016. *Questions*. Oxford University Press.

Jeroen Antonius Gerardus Groenendijk and Martin Johan Bastiaan Stokhof. 1984. *Studies on the Semantics of Questions and the Pragmatics of Answers*. Ph.D. thesis, Univ. Amsterdam.

Charles L. Hamblin. 1973. Questions in montague english. *Foundations of Language*, 10(1):41–53.

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

Lauri Karttunen. 1977. Syntax and semantics of questions. *Linguistics and philosophy*, 1(1):3–44.

Ai Kawazoe, Ribeka Tanaka, Koji Mineshima, and Daisuke Bekki. 2017. An inference problem set for evaluating semantic theories and semantic processing systems for japanese. In *New Frontiers in Artificial Intelligence: JSAI-isAI 2015 Workshops, LENLS, JURISIN, AAA, HAT-MASH, TSDAA, ASD-HR, and SKL, Kanagawa, Japan, November 16-18, 2015, Revised Selected Papers*, pages 58–65. Springer.

Manfred Krifka. 2001. For a structured meaning account of questions and answers. In *Audiatur Vox Sapientia. A Festschrift for Arnim von Stechow*, pages 287–320. Academie Verlag.

Manfred Krifka. 2003. Quantifiers in questions. *Korean Journal of English Language and Linguistics*, 3:499–526.

Yusuke Kubota and Robert D Levine. 2020. *Type-Logical Syntax*. MIT Press.

Tom Kwiatkowski, Luke Zettlemoyer, Sharon Goldwater, and Mark Steedman. 2011. Lexical generalization in ccg grammar induction for semantic parsing. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 1512–1523.

Mike Lewis and Mark Steedman. 2014. A* CCG parsing with a supertag-factored model. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 990–1000.

Koji Mineshima, Pascual Martínez-Gómez, Yusuke Miyao, and Daisuke Bekki. 2015. Higher-order logical inference with compositional semantics. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2055–2061, Lisbon, Portugal. Association for Computational Linguistics.

Rani Nelken and Chung-chieh Shan. 2004. A logic of interrogation should be internalized in a modal logic for knowledge. In *Semantics and Linguistic Theory*, volume 14, pages 197–211.

Rani Nelken and Chung-Chieh Shan. 2006. A modal interpretation of the logic of interrogation. *Journal of Logic, Language and Information*, 15:251–271.

Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know what you don't know: Unanswerable questions for SQuAD. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 784–789, Melbourne, Australia. Association for Computational Linguistics.

Siva Reddy, Mirella Lapata, and Mark Steedman. 2014. Large-scale semantic parsing without question-answer pairs. *Transactions of the Association for Computational Linguistics*, 2:377–392.

Mark Steedman. 1996. *Surface Structure and Interpretation*. MIT Press.

Mark Steedman. 2000. *The Syntactic Process*. MIT press.

Willemien Katrien Vermaat. 2005. *The Logic of Variation: A cross-linguistic account of wh-question formation*. Ph.D. thesis, Utrecht University.

Kazuki Watanabe, Koji Mineshima, and Daisuke Bekki. 2019. Questions in dependent type semantics. In *Proceedings of the Sixth Workshop on Natural Language and Computer Science*, pages 23–33, Gothenburg, Sweden. Association for Computational Linguistics.

Yimei Xiang. 2021. Binding without variables: Solving the under-generation problems. In *Semantics and Linguistic Theory*, volume 31, pages 001–020.

Masashi Yoshikawa, Hiroshi Noji, and Yuji Matsumoto. 2017. A* CCG parsing with a supertag and dependency factored model. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 277–287, Vancouver, Canada. Association for Computational Linguistics.