Dialogue & Discourse 13(2) (2022) 79–132 doi: 10.5210/dad.2022.203

Characterizing the Response Space of Questions: data and theory

Jonathan Ginzburg YONATAN.GINZBURG@U-PARIS.FR
Université Paris Cité, CNRS,
Laboratoire de Linguistique Formelle

Zulipiye Yusupujiang ZULPIYA127@HOTMAIL.COM
Université Paris Cité, CNRS,
Laboratoire de Linguistique Formelle

Chuyuan Li LISA27CHUYUAN@GMAIL.COM
Université de Lorraine, CNRS,
Inria, LORIA,

Kexin Ren KREN4@UALBERTA.CA
Université Paris Cité, CNRS,
Laboratoire de Linguistique Formelle

Aleksandra Kucharska ALE.KUCHARSKA96@GMAIL.COM
GSK (GlaxoSmithKline)

Paweł Łupkowski PAWEL.LUPKOWSKI@AMU.EDU.PL
Faculty of Psychology and Cognitive Science, Adam Mickiewicz University
Reasoning Research Group

Editor: Massimo Poesio

Submitted 12/2020; Accepted 10/2022; Published online 12/2022

Abstract

The main aim of this paper is to provide a characterization of the response space for questions using a taxonomy grounded in a dialogical formal semantics. As a starting point we take the typology for responses in the form of questions provided in Łupkowski and Ginzburg (2016). That work develops a wide coverage taxonomy for question/question sequences observable in corpora including the BNC, CHILDES, and BEE, as well as formal modeling of all the postulated classes. This paper extends that work to cover all types of responses to questions. We present the extended typology of responses to questions based on studies of the BNC, BEE, Maptask and CornellMovie corpora which include 607, 262, 460, and 911 question/response pairs respectively. We compare the data for English with data from Polish using the Spokes corpus (694 question/response pairs), providing detailed accounts of annotation reliability and disagreement analysis. We sketch how each class can be formalized using a dialogical semantics appropriate for dialogue management, concretely the framework of KoS (Ginzburg, 2012).

Keywords: question, responses, dialogue, corpus study

King Midas: What is the best thing for humans and the most choice worthy thing of all?' Silenos: Why are you forcing me to tell you humans what it would be better for you not to know? (Aristotle, *Eudemian Ethics* (Aristotle, 2012))

©2022 Jonathan Ginzburg, Zulipiye Yusupujiang,Chuyuan Li, Kexin Ren, Aleksandra Kucharska, Paweł Łupkowski
This is an open-access article distributed under the terms of a Creative Commons Attribution License (http://creativecommons.org/licenses/by/3.0/).

1. Introduction

There are various theories of what questions are (Groenendijk and Stokhof, 1997; Wiśniewski, 2015), and several computational theories of dialogue (Poesio and Rieser, 2010; Asher and Lascarides, 2003; Ginzburg, 2012), but no attempt yet at a comprehensive characterization of the response space of questions. Thus, our aim in this paper is to provide a characterization of the response space for questions using a taxonomy grounded in a dialogical formal semantics.[1]

This task, nonetheless, is of considerable theoretical and practical importance: it is an important ingredient in the design of dialogue systems, spoken or text–based; it provides benchmarks for dialogue/question theories, and of course is a component in explicating intelligence to pass the Turing test (see Turing, 1950).[2]

Łupkowski and Ginzburg (2013, 2016) tackled one part of this problem, offering an empirical and theoretical characterization of the range of *query* responses to a query (q-responses). Based on a detailed analysis of the British National Corpus and three other corpora, two task–oriented, (BEE (Rosé et al., 1999) and AmEx (Kowtko and Price, 1989)) and a sample from CHILDES (MacWhinney, 2000), they identified 7 classes of questions that a given query gives rise to; we refer to these classes as the L(upkowski)G(inzburg) classes of query responses. The study sample consisted of 1,466 query/query response pairs. As an outcome the following query responses taxonomy was obtained: (1) CR: clarification requests; (2) DP: dependent questions, i.e., cases where the answer to the initial question depends on the answer to a q-response; (3) MOTIV: questions about the underlying motivation behind the initial question; (4) NO ANSW: questions whose aim is to avoid answering the initial question; (5) FORM: questions which consider how to answer the initial question; (6) IND: questions which indirectly convey an answer, (7) IGNORE: responses ignoring the initial question, but addressing a shared situation—for more details see (Łupkowski and Ginzburg, 2016, p. 255). We take their work as a starting point and make the following hypothesis:

(1)(H) *Main hypothesis*: responses drawn from or concerning the LG query classes plus direct answerhood exhaust the response space of a query.

Specifically this amounts to the following general types of responses (we present the detailed taxonomy in Section 3).

1. Question–Specific:
   (a) Answerhood;
   (b) Dependent questions (A: Who should we invite? B: Who is in town?);
2. Clarification Responses.
3. Evasion responses:

[1]This paper is a substantially extended version of a paper that was presented at SigDial 2019 ("Characterizing the Response Space of Questions: a Corpus Study for English and Polish"). It includes a significantly broader review of the literature, the corpus study, manually annotated given the complexity of the categories, includes an additional corpus (the Cornell Movie corpus) and many more q/r pairs analyzed for English (1,235 vs. 2,240) and Polish (205 vs. 694); the discussion of annotation reliability is more extensive; the formal section has been rewritten and expanded considerably and the paper is also accompanied by two appendices covering formal background and the annotation guidelines.

[2]For the analysis of the Turing test as a question-response system see, e.g. (Łupkowski and Wiśniewski, 2011).

(a) Ignore (address the situation, but not the question);

(b) Change the topic ('Answer *my* question');

(c) Motive ('Why do you ask?');

(d) Difficult to provide a response.

The hypothesis has to be understood *relationally*—one is not really interested in the extension of the semantic entities (primarily propositions and questions) that can be given as responses. Rather, one is interested in the class each such entity is classified as since that is what determines the subsequent contextual evolution.

(2) I do not want to talk about that question. (Direct answer to *what do you not want to do?* Evasion answer to *Where were you last night?*).

We survey the existing literature in Section 2. Following this, we provide a description of the proposed taxonomy, in Section 3. In the sections that follow we proceed to test our main hypothesis using four corpora in English (BNC (Burnard, 2007), BEE (Rosé et al., 1999), HCRC MapTask (Anderson et al., 1991), CornellMovie (Danescu-Niculescu-Mizil and Lee, 2011)) and one corpus in Polish (Spokes; Pęzik 2014). Section 4 discusses respectively the corpora we used and data selected therefrom. Section 5 describes our annotation method.

The hypothesis achieves wide coverage, as we discuss in Section 6; in Section 7 we discuss in extensive detail the reliability of the results.

In Section 8 we consider the requirements on semantic frameworks for a formal characterization of the various classes of the taxonomy. We sketch an account of the different classes in the framework of KoS (Ginzburg, 2012), building on though departing in some respects from the account developed in (Łupkowski and Ginzburg, 2016). We point to problems other existing frameworks face in providing a comprehensive account. A concluding Section 9 outlines a variety of natural extensions to the work described here. There are two appendices: Appendix A offers basic notions from the type logical framework TTR (Cooper, 2012, 2023) used in the paper, whereas Appendix B provides the annotation guidelines.

2. Related work

As Enfield (2010, p. 2658) points out 'While the grammatical and information structural properties of questions have received widespread attention in linguistics literature, there has been relatively little attention paid to the *relationship between* questions and their responses.'

Let us start with Berninger and Garvey (1981) who introduce three terms to refer to a reaction to a question: (1) *response*, which is any verbal production emitted by a partner following a question; (2) *reply*, which is a response relevant to the question; and (3) *answer*—a reply that directly or indirectly provides the missing information. In what follows, the authors introduce their rich taxonomy of possible replies for children conversation in a nursery school. The taxonomy covers six categories: (1) Possible answers; (2) Indirect answers; (3) Confessions of ignorance; (4) Clarification questions; (5) Evasive replies; (6) Miscellaneous.

In particular, we find questions as a form of replying to questions among the proposed types (in the form of clarification questions). Further replies of this kind may be observed among the proposed sub-types of evasive replies (see 8 and 9—however, they are not as fine-grained as the LG typology of q-responses). These cover the following (see Berninger and Garvey, 1981, p. 407–408).

1. Selecting own reference in making an assertion:

   (3) X: Where the morrow's house?
   Y: Nope, well the morrow house has sniffles.

2. Selecting own reference in rejecting the presupposition of the question:

   (4) X: What's his name?
   Y: Um pretend that he didn't have a name.

3. Routinely associating question and answer form:

   (5) X: Why: (ellided: should I go to sleep).
   Y: Because.

4. Temporarily stalling in providing an answer, but acknowledging that question has been heard:

   (6) X: Now what do you want for dinner?
   Y: Well.
   X: Hot beef?
   Y: OK, hot beef.

5. Challenging questioner to supply answer:

   (7) X: What is it?
   Y: Guess.
   X: The lights?
   Y: Yes, it's a light yea I know.

6. Asking a related question other than for clarification purposes:

   (8) X: Where's my baby's food?
   Y: Are you ready for your baby's food?

7. Repeating question:

   (9) X: Where's Chrissy?
   Y: Where Chrissy?

8. Rejecting question as stated:

   (10) X: Do you hear the man that is with Lisa?
   Y: They're not with Lisa. I'm Lisa.

One may observe that the presented categories are co–extensive with the ones mentioned in the introduction to this paper. Possible and indirect answers are subsumed by the *question-specific: answerhood* category. Clarification questions correspond directly to the category of *Clarification responses*. And evasive replies and confessions of ignorance fall under our richer category dubbed *evasion responses*. Our proposed typology identifies also other types of question responses that are not tackled by Berninger's and Garvey's proposal.

In later work, an interesting typology of question responses was proposed as a result of an extensive 10-language comparative project on question–response sequences in ordinary conversation. The project was carried out from 2007 as the part of the Multimodal Interaction Project at the Max Planck Institute for Psycholinguistics—see an overview in (Stivers et al., 2010). The study adopted certain restrictions with respect to the questions which were taken into account. In order for a question-response pair to be coded the question had to be a formal question or a functional question. Questions seeking acknowledgment, offered in reported speech and requests for immediate physical action were not coded (Stivers and Enfield, 2010, p. 2621). The coding scheme for the response types presented in (Stivers and Enfield, 2010, p. 2624) is the following:

Non-response was coded if the person did nothing in response, directed his/her attention to another competing activity, or initiated a wholly unrelated sequence.

Non-answer response covers a verbal or visible response that failed to directly answer the question as put. This includes laughter, 'I don't know', initiation of repair (e.g., 'What?') or other inserted sequences, gestural responses such as shrugs that do not answer the question. Non-answer responses include also 'Maybe', 'Possibly' or responses that deal with the question indirectly (like e.g., A: 'Do you see Jack much?' B: 'He moved').

Answer Answers the question directly. Answers can be gestural (e.g., a head nod or shake) or verbal ('Uh huh', 'Yeah', or longer, more involved answers including partial repeats of the question to confirm or disconfirm).

Can't determine can't hear/see participants, etc.

As with the previous typology, one can observe that our categories of question response cover these discussed above. The types which are not covered (like parts of 'non-response' or 'can't determine categories') are a consequence of the set-up of the Multimodal Interaction Project, where annotators had video-taped conversations at their disposal. Our study is based on a wide range of already existing corpora (without access to video).

Another interesting issue concerns what constitutes the most frequent type of response. Berninger and Garvey (1981) observe that the vast majority of responses provided (for *polar* and for *Wh*-questions) were the possible *answers*. Other types were rare: 'The only other classes of replies that occurred with sizeable frequency were evasive replies and confessions of ignorance following *Wh*-questions and indirect answers following *yes/no* questions' (Berninger and Garvey, 1981, p. 410).

Analogous results are reported in Stivers and Robinson (2006) for the group of adult American English speakers. The corpus gathered for the analysis consisted of 260 instances of question sequences in a multi-party interaction (retrieved from video recordings of naturally occurring interactions). In this case the authors do not provide an extensive typology of replies as discussed

above, but focus only on answer / non-answer patterns. The conclusion of the study is that an answer is the alternative preferred over a non-answer (Stivers and Robinson, 2006, p. 371)—85% of the cases in the analyzed sample were answers. Stivers and Robinson provide several explanations for such a distribution. One is that the form of a non–answer supplying response turn reflects their ranking as *dispreferred* (they are frequently delayed both within and between turns, prefaced by filled pauses and discourse markers such as 'Well', and expanded with accounts—see Stivers and Robinson 2006, p. 372). Moreover, conversational participants typically treat a non-response as indicating disalignment, rather than indicating that no response will be forthcoming. Another reason for the obtained distribution, according to Stivers and Robinson, is that speakers perform interactional work to provide answers and despite the fact non-answers are a readily available alternative category of response—they 'struggle to receive and provide answers if at all possible' (Stivers and Robinson, 2006, p. 374). Stivers et al. (2010, p. 2616) point out that 'In English there is a strong normative order surrounding questions. In the first place, responses are normatively required [...], and answers are preferred over non-answer responses'. This claim is confirmed in the study of 350 questions drawn from spontaneous conversation in American English presented in (Stivers, 2010). The results are that 76% of responses were answers, only 19% were non-answers and 5% non-responses (Stivers, 2010, p. 2778). This is in line with previous results reported in (Stivers and Robinson, 2006) discussed above. Interestingly, Yoon (2010) reports results for Korean which though indicative of a similar pattern (Answer $>$ Non-Answer $>$ Non-response) indicate a markedly different distribution: of the sample of 326 questions-responses, 52% were answers, 33% non-answers and 15% non-responses (Yoon, 2010, p. 2790). In this study, the question sample was limited to questions that functionally sought information, confirmation or agreement (Yoon, 2010, p. 2783).

Enfield et al. (2019) present results of a fourteen-language (including e.g., English, Lao, Korean) study concerning the issue of how people answer polar questions. The data-set consisted of 172 videotaped interactions. The authors point out that they focus only on answers: "In our quantitative study of responses, we examine only confirming answers (rather than non-answers such as I don't know, I can't remember, or laughter; or disconfirming answers). This is because confirmations are more frequent than disconfirmations (...)" (Enfield et al., 2019, 288–289); it is worth noting that the non-answer examples acknowledged above are covered by our taxonomy. Enfield et al. (2019) conclude that the answers to polar question may be of two possible types: (i) interjection-type answers (such as 'uh-huh' or equivalents 'yes', 'mm', 'head nods', etc.)[3] and (ii) repetition-type answers. Wang (2020) uses the proposed taxonomy of polar-question answers in a study of Mandarin data, adding a 15th language to the already existing data.

Another notable source is Enfield (2010), who provides an analysis of questions and responses in Lao for a corpus of 351 questions drawn from 8 separate recordings. The results reported in this paper are interesting for the discussion of what counts as an answer to a question. The focus of the analysis is the structural fit between questions (wh–questions and polar ones) and their responses. The author offers the following hypothesis as to what answers to wh–questions are optimally coherent: '[the answer] should supply a referent of the relevant ontological category (i.e. a thing for a 'what' question, a person for a 'who' question, etc.).' (Enfield, 2010, p. 2661).

Green and Carberry (1999) provide useful insights into indirect answering. They study 25 dialogue examples originating in (Stenstrom, 1984), where 13% responses to polar questions were

[3]Further three sub-types of interjections answers (upgraded, downgraded, and acquiescent) were proposed in (Stivers, 2019).

indirect answers. On this basis one can highlight four possible reasons for using indirect answers (see Green and Carberry, 1999, p. 392).

1. To answer implicit wh-questions:

   (11) Q: Isn't your country seat there somewhere?
   R: [Yes/No].
   Stoke d'Abernon.

2. For social reasons:

   (12) Q: Did you go to his lectures?
   R: [Yes.]
   Oh he had a really caustic sense of humour actually.

3. To provide an explanation:

   (13) Q: And also did you find my blue and green striped tie?
   R: [No.]
   I haven't looked for it.

4. To provide clarification:

   (14) Q: I don't think you've been upstairs yet.
   R: [Yes, I have been upstairs.]
   Um only just to the loo.

The works discussed in this section indicate the need for a wider corpus study of the whole spectrum of responses to questions. These studies are limited in terms of the examples that were analyzed. They also impose certain limitations concerning the number of response categories to be identified. This is understandable, as their main aim was to explicate the answer/non-answer difference. We believe that an extensive corpus study should bring a fine grained characterization of the entire response space of questions. Moreover, we aim at providing an explicit dialogical semantics for each category in our corpus-based typology.

One should also acknowledge here the existence of various question answer typologies created within the field of Question Answering (QA). QA may be characterized as 'a sophisticated form of information retrieval (IR), in which the system processes questions queried in a natural language format and provides either the content containing the answer or the answer itself' (Shah et al., 2019, p. 611–612). Usually, such typologies are proposed for well-structured knowledge bases (or extracted with the use of various NLP methods from the unstructured data sets). What is different from our approach is that these typologies focus on the non-dialogical context of large amounts of texts. As such, QA addresses the interaction between computer systems and users (see Shah et al., 2019, p. 612). The resulting answer typology usually takes the form of an ontology of the related data on which a given QA system is to operate—an example of such an approach is presented in (Hovy et al., 2002).

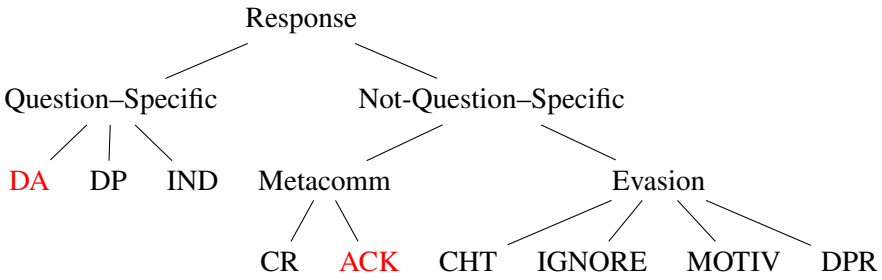


Figure 1: Proposed response space of questions

3. A taxonomy of responses to questions

Our taxonomy with its three main sub-partitions is displayed in Figure 1. The classes in red are those that were added by comparison with the query response taxonomy of Łupkowski and Ginzburg (2016).[4]

We start by the most general division of question responses to those that are specific to the question asked those that are not, as discussed in the introduction. In the question–specific class we distinguish direct from indirect answers and dependent questions.[5]

Direct answers (DA) provide an answer straightforwardly.[6,7] This is clearly visible in the following example—B is providing information required by A:

(15) A: Who is going to check that?

B: *Well I can check it.* [BNC: D97, L1226–1227]

For **indirect answers** (IND) one needs to infer an answer from the utterance.[8] This is exemplified in (16):

(16) A: What is it?

A: What's he done?

B: *Ehm, you know what I've said before, eh, eh you'll get <unclear>.* [BNC: KD5, L175–L177]

[4]For an explicit presentation of the taxonomy *sans* answers, see (Łupkowski and Ginzburg, 2016, p. 256).

[5]An anonymous reviewer for *Dialogue and Discourse* suggests that, directness may be understood as a separate dimension, which is independent from the others. They suggest that any type of response may be presented in direct or indirect manner (not just answers). This is a hypothesis we think is worth testing, though we do not do so in the current paper.

[6]We give a more explicit characterization of answerhood in Section 8; for a thorough, historically based discussion see (Wiśniewski, 2015).

[7] For the direct answers category we allow for additional sub-categories, which we did not use in the annotation, but which we return to discuss briefly in Section 8. These include: (1) no/yes answer to polar questions; (2) simple answers to wh-questions; (3) partial polar answers; (4) partial wh-question answers.

[8]As with the direct answers category, it is also apt to use the following sub-categories of indirect answers: (i) indirect answers addressing a wh-question; (2) q-widening INDs (over-informative answer to a polar question, addressing a more general wh-question).

In (16) A needs to infer the answer to his/her questions from B's suggestion that this issue has been addressed before. One also encounters IND being a question-response, as in (17), which is rhetorical and in this sense does not need to be answered and **indirectly provides an answer** to the initial question (q1).

(17) A: Are you Gemini?

B: *Well if I'm two days away from your, what do you think?* [BNC: KPA, L3603–L3604]

Dependent questions (DP) constitute the case where the answer to the initial question (q1) depends on the answer to the query-response (q2), as in:

(18) a. A: q_1

B: q_2

$\mapsto q_1$ depends on q_2

b. A: Do you want me to <pause> push it round?

B: *Is it really disturbing you?*
[*cf. Whether I want you to push it around depends on whether it really disturbs you.*] [BNC: FM1, 679–680]

The other two remaining super-categories reuse the classes proposed in (Łupkowski and Ginzburg, 2013, 2016) with some minor renaming. We start with the *metacommunicative* class, involving Clarification responses and acknowledgments.

Clarification responses (CR) address something that was not completely understood in the initial question (q1)[9], like:

(19) A: Why are you in?

B: *What?* [BNC: KPT, 469–470]

Some significant consequences this class has for contextual composition is discussed in Section 8.2.

Acknowledgment (ACK)—a speaker acknowledges that s(he) has heard and understood the question, e.g. *mhm*, *aha* etc.[10]

(20) a.

LEELOO: Do you know how we say 'make love'?

KORBEN: *Uh...*

LEELOO: ...Hoppi-hoppa [Cornell Movie, 5963-5965]

[9]This class contains intended content questions, repetition requests and relevance clarifications—for detailed discussion see e.g. (Purver, 2006) or (Ginzburg, 2012).

[10]Acknowledgments are much rarer after questions than after assertoric moves, often communicating, as in two of the examples here, hesitation as to how to answer the question; a finer grained scheme might distinguish such cases from "pure" backchannels with a continuative import.

b.

A: Who's it for? <pause>

B: *Er*

A: Private job or [BNC: KD3, 2063–2065]

c.

A: What's that called, the centre line of the earth?

B: *Mm* [BNC: F72, 62–63]

Moving on to evasive question-responses, we mention first the type which addresses the **motivation underlying asking q1** (MOTIV). Whether an answer to q1 will be provided depends on a satisfactory answer to q2, as in (21a); (21b) is an instance where the responder offers an answer negatively resolving the motivation issue:

(21) a. A: What's the matter?

B: *Why?* [BNC: HMD, 470–471]

b.

REPORTER: Who did you back prime minister?

THERESA MAY: *As I said last week none of your business.* [The Guardian, May 2019]

A related class, which was subsumed within MOTIV in (Łupkowski and Ginzburg, 2016),[11] but which we separate away here involves cases where the speaker states that it is **difficult to provide an answer** (DPR), points at a different information source, etc. or the speaker states that s(he) **does not know the answer**.

(22) TUTOR: Why?

STUDENT: *i'm not exactly sure.* [BEE: log-stud29]

(23) A: When's the first consignment of Scottish tapes?

B: *Erm <pause> don't know.* [BNC: FM2, 1061–1062]

Another type of evasive question-response is **change–the–topic (CHT)**. Instead of answering q1, the agent directly provides q2 and attempts to turn the table on the original querier. The original querier is pressured to answer q2 and put q1 aside, as exemplified in (24a) and most explicitly in (24b).[12]

[11] This subclass was insubstantial when solely query responses are considered.

[12] These can occur in text as well:

(i) So, in answer to the question: Is Jeremy Corbyn an anti-Semite? My response would be that that's the wrong question. The right questions to ask are: Has he facilitated and amplified expressions of anti-Semitism? Has he been consistently reluctant to acknowledge expressions of anti-Semitism unless they come from white supremacists and neo-Nazis? Will his actions facilitate the institutionalization of anti-Semitism among other progressives? Sadly, my answer to all of these is an unequivocal yes. [D. Lipstadt, *Antisemitism: Here and Now*, p. 67]

(24) a.

A: What we doing in that?

A: Er

B: *So did this woman ask you about why you've had so many Fridays off?* [BNC: KNY, 1005–1007]

b.

BBC INTERVIEWER: How did Singapore handle the pandemic so well?

SINGAPORE HEALTH OFFICIAL: The question should be 'How did UK not handle it so well?'.

BBC INTERVIEWER: What do you mean?

SINGAPORE HEALTH OFFICIAL: We followed 'UK Pandemic Response Protocol', the UK did not! [Twitter 24 May 2021]

(25) provide examples of propositional CHT, where the response addresses a distinct issue, thereby indicating that this latter is the topic the responder wishes to discuss and not the initial issue:

(25) a.

A: What's dolly's name?

B: *It's raining.* [BNC: KD4, 110–111]

b.

KAT: You're amazingly self-assured. Has anyone ever told you that?

PATRICK: *Go to the prom with me!* [Cornell Movie Corpus, m6, 839–840]

An **IGNORE** type of query-response appears when q2 relates to the situation described by q1 but not directly to the initial question. This can be observed in (26). A and B are playing Monopoly. A asks a question, which is ignored by B. It is not that B does not wish to answer A's question and therefore asks q2. Rather, B ignores q1 and asks a question related to the situation (in this case, the board game).

(26) A: I've got Mayfair <pause> Piccadilly, Fleet Street and Regent Street, but I never got a set did I?

B: *Mum, how much, how much do you want for Fleet Street?* [BNC: KCH, 1503–1504]

See also the following examples:

(27) A: Just one car is it there?

B: *Why is there no parking there?* [BNC: KP1, 7882–7883]

Similar examples emerge with propositional responses, as evinced in (28) and (29):

(28) A: So does that mean that the ammeter is not part of the series, just hooked up after to the tabs?

B: *Let's take a step back.* [BEE, log-stud23]

(29)

DINO VELVET : Mister Welles . . . would you be so kind as to remove any firearms from your person?

WELLES: What are you... ?

DINO VELVET : *Take out your gun!* [Cornell Movie Corpus, 6840–6842]

4. Corpus data used for the study

In order to test our main hypothesis, we used corpora from two languages: English and Polish.

4.1 English: BNC, BEE, MapTask, CornellMovie

The data for English comes from the BNC (Burnard, 2007), BEE (Rosé et al., 1999), MapTask (Anderson et al., 1991; Skantze et al., 2006) and the CornellMovie corpora (Danescu-Niculescu-Mizil and Lee, 2011). Although both self–answering and multiparty turns figured in the initial development stage of the taxonomy, we restricted attention to two–person dialogue in the study reported here. 607 Q-R turns were taken from the BNC, 262 Q-R turns from BEE, 460 Q-R turns from the MapTask, and 911 Q-R turns from the CornellMovie corpus. The BNC data covers mainly free conversations: initially 864 Q-R pairs from BNC were annotated, but after elimination of multi-party segments, 607 Q-R pairs were retained. As for BEE, 37 undergraduate students with little background in electricity or electronics participated in conversations with a tutor. We randomly selected the students' numbers (23, 25, 27, 28, 29, and 31) and annotated the dialogues generated between those students and the tutor. In this way, we obtained 262 Q-R pairs. The MapTask consists of dialogues recorded for a route following task in which one participant directs a second participant along a route in a map, though the route giver and route follower maps are not identical. 297 of the 460 MapTask Q-R pairs are from the HCRC MapTask corpus (Anderson et al., 1991), whereas 163 of them are from the Higgins pedestrian navigation and guiding project (Skantze et al., 2006). The HCRC Map task corpus contains 128 dialogues, 64 of which involve eye contact between participants, while the remaining 64 dialogues involved no eye contact. In this study, we chose 28 dialogues, of which 14 with eye contact and 14 without eye contact. The filenames of these dialogues were selected randomly. We annotated all Q-R pairs occurring in each dialogue and obtained 297 Q-R annotated pairs after eliminating cases involving self–answering, incomplete questions, and overlapping. As for the Q-R turns from the Higgins project, we annotated six folders (files no.1 -no.6) and in each of them, there are 4 or 5 different dialogue files. As a result, we also annotated 28 dialogues and obtained 163 Q-R turns. The CornellMovie corpus is a collection of fictional conversations extracted from raw movie scripts. We annotated all available two-person dialogues from the first 8 movies listed in the corpus, ranging from the movie ID *m0* to *m7*, thereby obtaining 911 Q-R pairs. This covers various genres such as comedy, romance, adventure, biography, history, action, crime, science fiction, thriller, fantasy, and horror.

Table 1: Summary of the corpus data used for the study

Corpus	Q-R pairs	Domain
BNC	607	free conversations
BEE	262	tutorial dialogues
MapTask	460	cooperative task
CornellMovie	911	scripted conversations
Spokes	694	free conversations
Total	2,934	

4.2 Polish: the Spokes Corpus

The data used for this study were retrieved from the Spokes corpus. The corpus currently contains 247,580 utterances (2,319,291 words) in transcriptions of spontaneous conversations. For the purposes of this paper, two studies were conducted (with two different sets of annotators). For the first study, we selected four files from the corpus (10,244 words). For the second study, 21 files were selected (86,052 words). The files cover casual conversations concerning, e.g., youth, TV shows, children, wine, or travel plans. Within each file, the question-response pairs (Q-R) were selected manually. In total, we obtained 694 Q-R pairs for two studies.[13]

5. Annotation method

For the annotation, all the question-response pairs were supplemented with a full context. The guideline for annotators contained explanations of all the classes and examples for each category. Moreover, the OTHER category was included. The complete annotation guidelines are presented in Appendix B of this paper.

English data annotation: The 607 BNC Q-R turns used in this study were randomly extracted from the British National Corpus (BNC) and manually annotated by one English L1 speaker and two English L2 speakers who have masters degrees in Linguistics and underwent several training sessions with one of the authors, a native speaker of English with significant experience in dialogue annotation.

Among the 607 Q-R turns, 334 of them were annotated by the first and second annotators, whereas the remaining 273 Q-R turns were annotated by the first and the third annotators. Therefore, an inter-annotator study was conducted in two groups: first vs. second annotators, and first vs. third annotators.

Polish data annotation: The first sample of 205 Q-Rs was annotated by the main annotator and two other annotators (one of whom has previous experience in corpus data annotation, all annotators were Polish native speakers). The annotators received the annotation guidelines and underwent a short training phase based on selected examples. The second sample of 489 Q-Rs was annotated by the main annotator and two other annotators who are different from that of the first sample (the main annotator remained the same as in the first sample, all annotators were Polish native speakers). As

[13] Given that Spokes is the sole source we had for Polish, we did not restrict attention to two-person dialogue, given that this would have significantly reduced our data set.

in the previous case, annotators received the annotation guidelines and underwent a short training based on selected examples.

6. Results

The detailed results of the annotation are presented in Figure 2. We discuss the annotation reliability in Section 7. We also provide additional data for this paper (covering annotated Q-R pairs and disagreement cases) which are hosted on the OSF web-page (`https://osf.io/mq6r7/`).

6.1 English

In all four cases, the OTHER class is less than 0.5%, hence coverage is above 99.5%. The most frequent response classes in all four corpora are direct answers; the second most frequent class in the BNC is Difficult to provide an Answer (DPR=7.91%), while in CornellMovie, the next biggest is indirect answers (IND=18.33%), whereas for the MapTask and BEE these are IGNORE (6.09% and 3.82% respectively).

6.2 Polish

The two most frequent classes of responses for Spokes are answers: direct ones (DA=64.27%) and—much smaller—indirect ones (IND=10.66%). The next two most frequent classes are DPR (stating that a person does not know the answer to the question, or it is difficult to provide one, DPR=7.78%) and utterances ignoring the question asked (questions and declaratives, IGNORE=6.92%).

6.3 Discussion

When comparing results for English and Polish, it is apparent that the largest category is direct answers (DA). Also, indirect answers constitute a large group among recognized responses types in both languages. This result is in line with the findings reported by Stivers and Robinson (2006) and Yoon (2010)—summarized in Section 2.

As might be expected given the previous results presented in Łupkowski and Ginzburg (2016), the most frequent question-response for English and Polish data is the clarification request. What is interesting is the relatively high number of *ignoring* responses observed for English and Polish. In (Łupkowski and Ginzburg, 2016) we analyzed only question-responses and this type was observed rarely (0.57% for N=1,051 for BNC). This time, IGNORE has been used also to classify declaratives, which may explain the higher number observed; we discuss a possible semantic explanation for this in Section 8, where we suggest that it is in some sense only "weakly evasive". Other evasive responses (relatively) frequent in both languages are CHT and DPR.

We can also make some comments concerning cross-corpus differences. As we already mentioned in Section 4, our BNC and Spokes data cover mainly free conversations, while BEE and MapTask contain task-oriented dialogues. One might expect differences between these dialogue genres. These expectations are indeed fulfilled: the MapTask and BEE corpora have the highest number of direct answers in our study sample (80.87% and 88.93% respectively). In contrast, for the BNC and Spokes corpora these numbers are substantially lower (respectively 69.36% and 64.27%). When it comes to clarification responses, we observe that the numbers are lower for task-oriented corpora than for the BNC and Spokes (this is in line with our previous results for BNC and BEE, reported in (Łupkowski and Ginzburg, 2016, p. 256–257)). We also observe that, for the

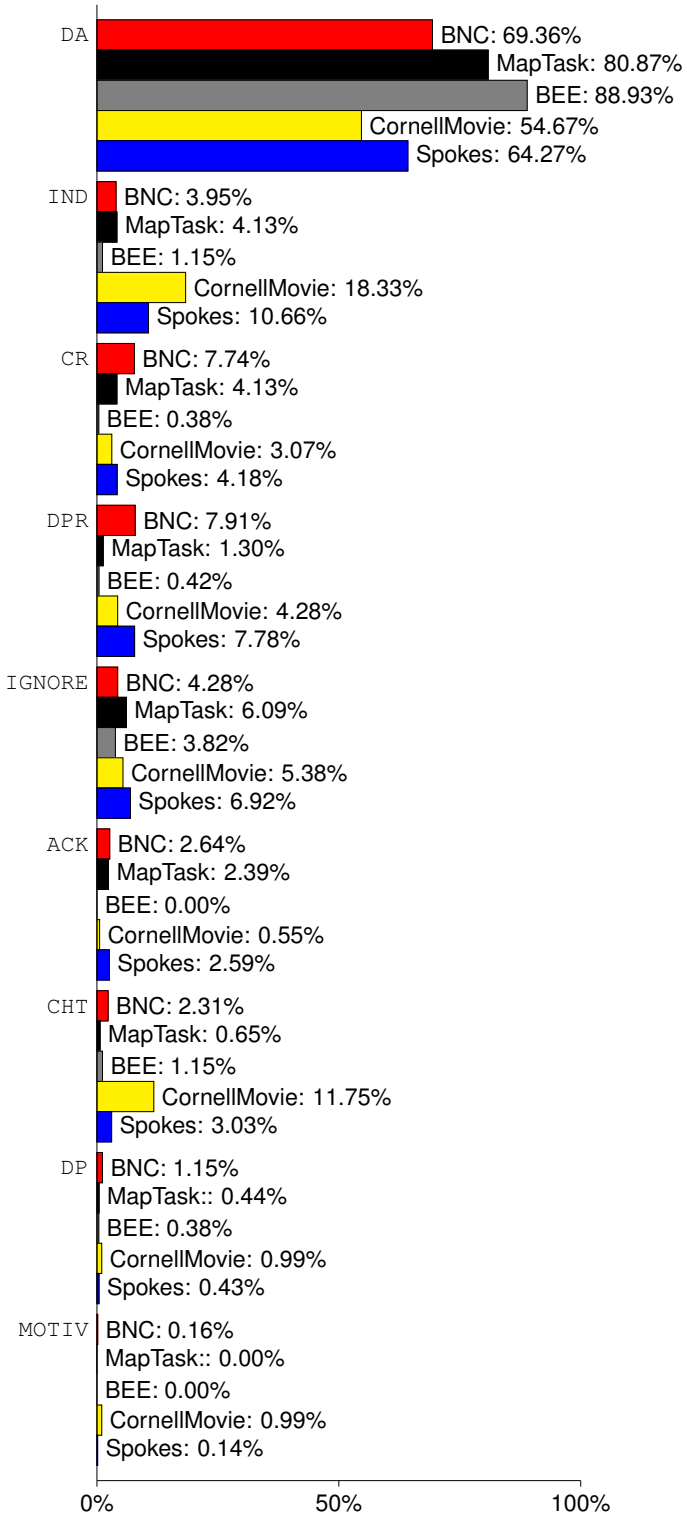


Figure 2: Response types frequency (BNC, n=607; BEE, n=262; MapTask, n=460; CornellMovie, n=911; Spokes, n=694)

evasive response types discussed above, the tendency is analogous, i.e., we observed lower numbers for task-oriented dialogues than for free conversations. For the CornellMovie corpus, which is a collection of fictional, scripted conversations extracted from raw movie scripts, we observe tendencies akin to the BNC and Spokes. This is more or less expected, as elite scriptwriters aim for and succeed in mimicking natural conversation. One notable exception is the CHT response category (11.75% vs 2.31%–3.03%). One may hypothesize that such an evasive response is especially useful for movie dialogue writers—however, this observation needs further investigation.

7. Annotation reliability

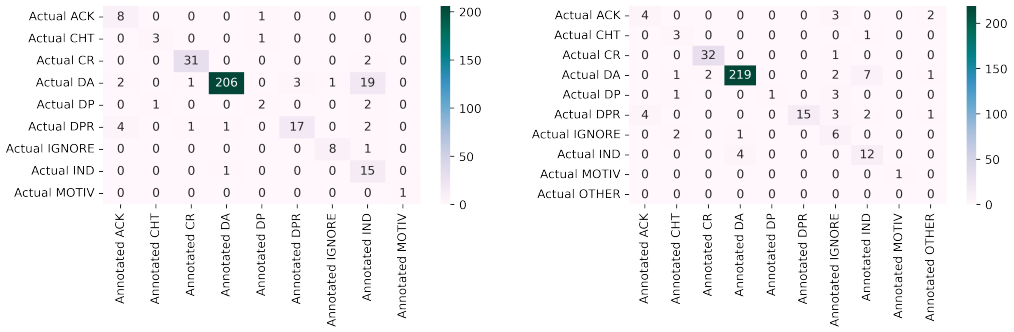
7.1 Inter-annotator study

We conducted the following inter-annotator reliability study on the English BNC and Polish Spokes corpora, as they were double annotated by multiple annotators. However, other English corpora such as BEE, MapTask, and CornellMovie were annotated only once.

English The reliability of the annotation was evaluated using the κ (Carletta, 1996) and α (Krippendorff, 2011) coefficients. We used the *Scikit-learn* (Pedregosa et al., 2011) data mining and data analysis tool in Python with its *sklearn.metrics* package for calculating Cohen's kappa, and also used the Python implementation *Krippendorff*[14] for the calculation of Krippendorff's alpha. In this case, Cohen's Kappa for the first and second annotators is 0.7053 (substantial), whereas for the first and third annotators it is 0.6430. Krippendorff's alpha for the first group is 0.7022, while 0.6373 for the second group. All disagreements were then discussed in detail by one of the annotators and the aforementioned author and resolved. As a result, we obtained a gold standard for this BNC annotation task. In addition, as seen in Figure 3 and Figure 4, we created a confusion matrix for each of these three annotators by comparing their annotations with the gold standard. We also calculated precision, recall, and F-1 measures of each class for all three annotators as shown in Table 2. All were calculated by using the data analysis tool *Scikit-learn* in Python with its *sklearn.metrics* package.

We can learn the annotation performance of each annotator by investigating the results shown in the confusion matrices in Figure 3 and Figure 4, as well as from the precision, recall, and F-1 scores reported in Table 2. For the largest categories, on the whole, DA and CR were easy to annotate; IND was more tricky. In more detail: for the first group annotation in English, there are 334 annotated Q-R pairs in total, and among them 232 are DA, 33 are CR, and 16 are IND. The first annotator correctly annotated 206 of DA as DA but misannotated 19 of them as IND, 2 as ACK, one as CR, 3 as DPR, and one as IGNORE. Therefore, the first annotator obtained a precision of 0.99, recall 0.89, and the F-1 score of 0.94 for the response type DA. The second annotator on the other hand, correctly annotated 219 out of 232 actual DA as DA but misannotated 2 of them as CR, 1 as CHT, 7 as IND, 2 as IGNORE, and 1 as OTHER. Therefore, the second annotator gained a precision of 0.98, recall 0.94, and the F-1 score 0.96 for the response type DA. As for the response type CR, the first and second annotators obtained a recall score, 0.94 and 0.97 respectively. That is, the first annotator correctly identified 31 out of 33 CR, and only misclassified 2 as IND. The second annotator identified all 32 CR correctly, and only misclassified one as IGNORE. The precision and F-1 score of CR for the first annotator is 0.94, and 0.94 and 0.96 for the second annotator. Regarding the response type IND, the first annotator correctly annotated 15 out of 16 as IND but misclassified 1

[14] `https://pypi.org/project/krippendorff/`



(a) (b)

Figure 3: **English First group** Confusion Matrices: (a) First Annotator (b) Second Annotator

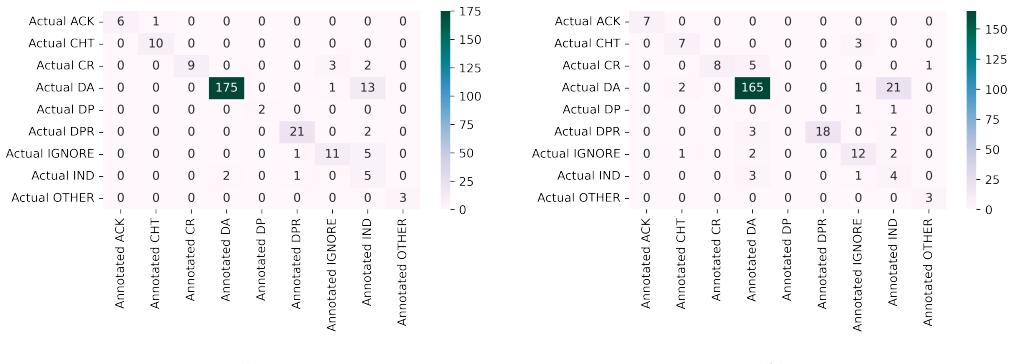
of them as DA. The precision, recall, and F-1 score of IND for the first annotator are 0.37, 0.94, and 0.53 respectively. The second annotator correctly identified 12 out of 16 as IND but misclassified 4 of them as DA. The precision, recall, and F-1 score of IND for the second annotator are 0.55, 0.75, and 0.63 respectively.

As shown in the Figure 4 and Table 2, in the second group of annotation for English, there are 273 annotated Q-R pairs in total, and among them, there are 189 DA, 8 IND, 14 CR, 23 DPR, and 17 IGNORE. The first annotator correctly annotated 175 DA as DA but misclassified 13 as IND, and one as IGNORE. The precision, recall, F-1 scores of DA for the first annotator are 0.99, 0.93, and 0.96 respectively. The third annotator correctly identified 165 out of 189 DA but misclassified 21 of them as IND, 1 as IGNORE, and 2 as CHT. The third annotator obtained a precision score of 0.93, recall 0.87, and F-1 score of 0.90. As for the response type CR, the first annotator correctly annotated 9 cases, but misidentified 3 as IGNORE, and 2 as IND. The precision, recall, and F-score for CR are 1.00, 0.64, and 0.78 respectively for the first annotator. The third annotator also performed similarly in the classification of CR, and she correctly annotated 8 out of 14 cases, but missclassified 5 as DA, and 1 as OTHER. The third annotator obtained similar performance scores as the first annotator, which are 1.00, 0.57, and 0.73 for precision, recall, and F-1 scores respectively. As to the response type IND, 5 out of 8 cases were annotated correctly by the first annotator. However, there are 2 cases misclassified as DA, one as DPR. The precision, recall, and F-1 score obtained by the first annotator are 0.19, 0.62, and 0.29 respectively. The annotation of IND also caused more difficulties to the third annotator. She correctly identified 4 out of 8 IND cases but misclassified 3 as DA, and 1 as IGNORE. The performance scores of the third annotator for the response type IND are 0.13, 0.50, and 0.21 respectively for precision, recall, and F-1 score. The annotation performance of both annotators on the response types IGNORE are similar. However, the F-1 scores of CHT are 0.95 and 0.70 for the first and the third annotator respectively, 1.00 and 0.00 for the response type DP. In addition, the annotators' performance of the second group is better than the first group in terms of the annotation of response classes DPR and ACK.

Polish The reliability of the annotation for Polish was also evaluated using the κ (Carletta, 1996) and α (Krippendorff, 2011) coefficients. As mentioned above, the main annotator was the same person in both samples. However, other annotators were different in these two annotation groups. The reported values were calculated using the same method and tools as for English. For the first

Table 2: Detailed Annotation Report for **English** Annotators

Annotator	Classes	Precision	Recall	F1-score	Frequency
First Group First Annotator	DA	0.99	0.89	0.94	232
	IND	0.37	0.94	0.53	16
	CR	0.94	0.94	0.94	33
	DPR	0.85	0.68	0.76	25
	IGNORE	0.89	0.89	0.89	9
	ACK	0.57	0.89	0.70	9
	CHT	0.75	0.75	0.75	4
	DP	0.50	0.40	0.44	5
	MOTIV	1.00	1.00	1.00	1
	accuracy			0.87	334
	macro avg.	0.76	0.82	0.77	334
	weighted avg.	0.92	0.87	0.89	334
First Group Second Annotator	DA	0.98	0.94	0.96	232
	IND	0.55	0.75	0.63	16
	CR	0.94	0.97	0.96	33
	DPR	1.00	0.60	0.75	25
	IGNORE	0.33	0.67	0.44	9
	ACK	0.50	0.44	0.47	9
	CHT	0.43	0.75	0.55	4
	DP	1.00	0.20	0.33	5
	MOTIV	1.00	1.00	1.00	1
	OTHER	0.00	1.00	0.00	0
	accuracy			0.88	334
	macro avg.	0.67	0.73	0.61	334
	weighted avg.	0.92	0.88	0.89	334
Second Group First Annotator	DA	0.99	0.93	0.96	189
	IND	0.19	0.62	0.29	8
	CR	1.00	0.64	0.78	14
	DPR	0.91	0.91	0.91	23
	IGNORE	0.73	0.65	0.69	17
	ACK	1.00	0.86	0.92	7
	CHT	0.91	1.00	0.95	10
	DP	1.00	1.00	1.00	2
	OTHER	1.00	1.00	1.00	3
	accuracy			0.89	273
	macro avg.	0.86	0.85	0.83	273
	weighted avg.	0.94	0.89	0.91	273
Second Group Third Annotator	DA	0.93	0.87	0.90	189
	IND	0.13	0.50	0.21	8
	CR	1.00	0.57	0.73	14
	DPR	1.00	0.78	0.88	23
	IGNORE	0.67	0.71	0.69	17
	ACK	1.00	1.00	1.00	7
	CHT	0.70	0.70	0.70	10
	DP	1.00	0.00	0.00	2
	OTHER	0.75	1.00	0.86	3
	accuracy			0.82	273
	macro avg.	0.80	0.68	0.66	273
	weighted avg.	0.89	0.82	0.84	273



(a) (b)

Figure 4: **English Second group** Confusion Matrices: (a) First Annotator (b) Third Annotator

sample, the best inter-annotator κ and α scores were achieved by the second and the main annotators, 0.6579 and 0.6582 respectively. While for the second sample, we observed the highest inter-annotator agreements between the first and the main annotators, which are 0.5467 and 0.5466 for κ and α, as shown in Table 3. All disagreements were discussed in detail by the main annotator and resolved. In addition, we also used the data analysis tool *Scikit-learn* in Python with its *sklearn.metrics* package to created a confusion matrix for each of five annotators by comparing their annotation with the gold standard, as well as calculated the precision, recall, and F-1 measures of each response type. The annotation performance of each Polish annotator is presented in detail in Table 4, Table 5, and Figure 5. The frequency of each response type for the **Polish first group** annotation is DA:107, IND:29, CR:9, DPR:22, IGNORE:23, ACK:3, CHT:11, DP:1, and MOTIV: 0. Regarding the most frequent response type DA, the first annotator correctly annotated 87 out of 107 DA cases but misclassified 13 as IND, 3 as IGNORE, 2 as DPR, 1 as CR, and 1 as OTHER. As a result, the first annotator obtained a precision of 0.94, recall 0.81, and F-1 score of 0.87. The second annotator correctly identified 6 more DA cases than the first annotator but also misannotated 5 as IND, 3 as CHT, 1 as CR, 2 as IGNORE, and 1 as DPR. The performance scores are also very close to those of the first annotator, which are 0.94, 0.89, and 0.91 for precision, recall, and F-1 score respectively. As to the response type IND, the first annotator correctly annotated 21 out of 29 IND cases but misclassified 3 as DA, and other 5 as CHT, CR, IGNORE, MOTIV, and OTHER respectively. The precision, recall, and F-1 score of IND for the first annotator are 0.48, 0.72, and 0.58. The second annotator correctly annotated 18 out of 29 IND cases but misclassified 5 as IGNORE, 2 as CHT, and other 4 as DA, DP, DPR, and OTHER respectively. The precision, recall, and F-1 scores for the second annotator are 0.75, 0.62, and 0.68. Regarding the response type CR, the first annotator successfully identified 5 cases, whereas the second annotator identified 7. The F-1 scores for the first and the second annotator are 0.59 and 0.82 respectively. As for IGNORE, the first annotator correctly identified only 12 cases out of 23, whereas the second annotator correctly classified 21. The F-1 scores of the response type IGNORE for the first and the second annotator are 0.53 and 0.78. Comparing the F-1 scores for each response type, we learned that the second annotator performed better than the first annotator in general.

When it comes to the **Polish second group** annotation, there are 489 annotated Q-R pairs in this sample. The frequencies of each response type are DA:339, IND:45, CR: 20, DPR:32, IGNORE:25, ACK:15, CHT:10, DP:2, and MOTIV:1. As for the response type DA, the first annotator correctly

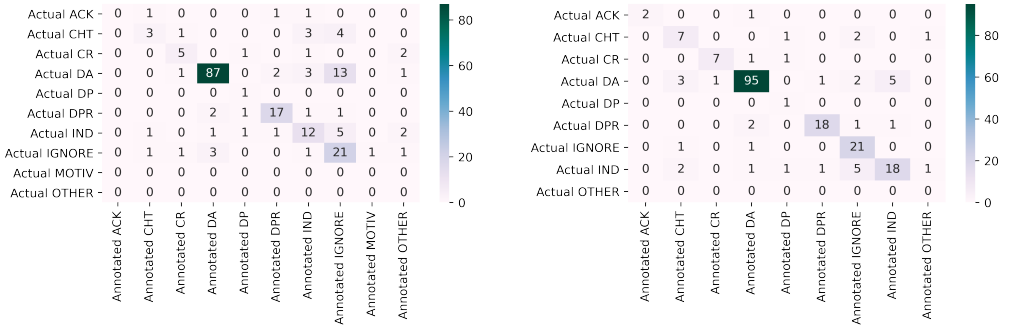
annotated 328 out of 339 DA cases but misclassified 4 as DPR, 2 as CHT, 2 as CR, and the remaining 3 as DP, IND, and IGNORE. The precision, recall, and F-1 score for the first annotator are 0.93, 0.97, and 0.95 respectively. The second annotator, on the other hand, correctly identified 266 out of 339 DA cases. The second annotator misclassified 33 DA cases as IND, 13 as DPR, 14 as IGNORE, 10 as CHT, 2 as ACK, and 1 as CR. As a result, the second annotator obtained a precision of 0.99, recall 0.78, and F-1 score of 0.88. Regarding the response type IND, the first annotator correctly annotated 34 out of 45 IND cases but misclassified 10 as DA, 1 as CH, and obtained a precision of 0.85, recall 0.76, and F-1 score of 0.80. The second annotator correctly identified 38 out of 45 IND cases, but misannotated 7 as IGNORE. The precision, recall, and F-1 score for the second annotator are 0.49, 0.84, and 0.62 respectively. As to the response type CR, the first annotator correctly annotated 9 out of 20 CR cases but failed to identify the remaining 11 cases. The precision, recall, and F-1 scores for the first annotator are 0.82, 0.45, and 0.58. The second annotator correctly annotated 11 out of 20 CR cases but misclassified 2 as DA, 3 as IND, and 4 as IGNORE. The precision, recall, and F-1 scores for the second annotator are 0.92, 0.55, and 0.69. Regarding the response type IGNORE, the first annotator correctly identified 14 out of 25 IGNORE cases but misclassified 6 as CHT, 3 as DA, and 2 as IND. The precision, recall, and F-1 scores are 0.82, 0.56, and 0.67 for the first annotator. Whereas the second annotator correctly annotated 23 out of 25 IGNORE cases and only misclassified 2 of them as CHT. Even though the second annotator obtained a high recall of 0.92, he has a low precision and F-1 score, which are 0.43 and 0.58 respectively. In addition, the first and the second annotators performed similarly on the annotation of the response types DPR, ACK, CHT, and MOTIV. However, as for DP, the second annotator obtained 1.00 for all the precision, recall, and F-1 scores, and the first annotator obtained 0.40, 1.00, and 0.57 respectively.

As for the performance of the main annotator in both groups of annotation samples, he outperformed all the other annotators in most of the cases. However, in the first group of samples, the main annotator failed to correctly capture the response type DP, which has only one case in this sample. As for all other response types, he obtained very high F-1 scores, which are above 0.90 in most cases, and 0.83 and 0.87 for IND and CHT respectively. While in the second group of samples, he did not perform as well as the other annotators regarding ACK. He only obtained an F-1 score of 0.55, while the other two obtained 0.97 and 0.94 respectively. In addition, he also only got an F-1 score of 0.50 for the response type DP. What's more, the first annotator outperformed the main annotator also on the annotation of IND, where the first annotator obtained an F-1 score of 0.80, whereas it is 0.73 for the main annotator.

Table 3: Polish inter-annotator agreement

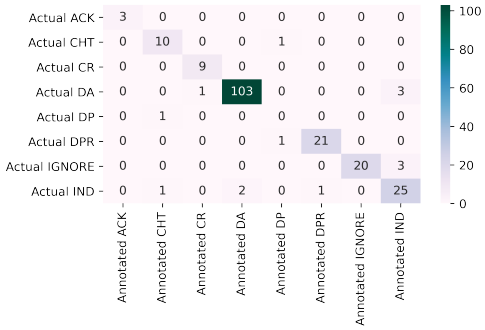
Annotation Group	Annotators	Cohen's kappa	Krippendorff's alpha
First Group	First Annotator vs. Second Annotator	0.4588	0.4574
	First Annotator vs. Main Annotator	0.5121	0.5117
	Second Annotator vs. Main Annotator	0.6579	0.6582
Second Group	First Annotator vs. Second Annotator	0.5414	0.5334
	First Annotator vs. Main Annotator	0.5467	0.5466
	Second Annotator vs. Main Annotator	0.4738	0.4648

Annotation reliability on the subsets of Taxonomy

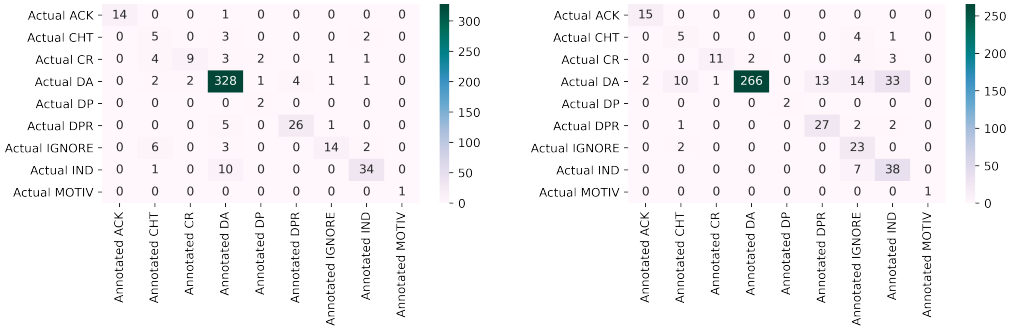


(a)

(b)

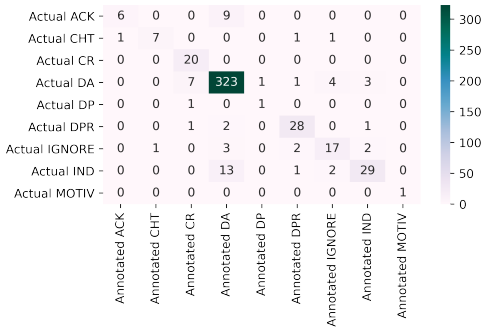


(c)



(d)

(e)



(f)

Figure 5: **Polish** Confusion Matrices: (a)-(c) First Group Annotators; (d)-(f) Second Group Annotators

Table 4: Detailed Annotation Report for **Polish First Group** Annotators

Annotator	**Classes**	**Precision**	**Recall**	**F1-score**	**Frequency**
First Group First Annotator	DA	0.94	0.81	0.87	107
	IND	0.48	0.72	0.58	29
	CR	0.62	0.56	0.59	9
	DPR	0.81	0.77	0.79	22
	IGNORE	0.55	0.52	0.53	23
	ACK	1.00	0.00	0.00	3
	CHT	0.50	0.27	0.35	11
	DP	0.25	1.00	0.40	1
	MOTIV	0.00	1.00	0.00	0
	OTHER	0.00	1.00	0.00	0
	accuracy			0.71	205
	macro avg.	0.51	0.67	0.41	205
	weighted avg.	0.77	0.71	0.73	205
First Group Second Annotator	DA	0.94	0.89	0.91	107
	IND	0.75	0.62	0.68	29
	CR	0.88	0.78	0.82	9
	DPR	0.90	0.82	0.86	22
	IGNORE	0.68	0.91	0.78	23
	ACK	1.00	0.67	0.80	3
	CHT	0.54	0.64	0.58	11
	DP	0.25	1.00	0.40	1
	OTHER	0.00	1.00	0.00	0
	accuracy			0.82	205
	macro avg.	0.66	0.81	0.65	205
	weighted avg.	0.85	0.82	0.83	205
Main Annotator	DA	0.98	0.96	0.97	107
	IND	0.81	0.86	0.83	29
	CR	0.90	1.00	0.95	9
	DPR	0.95	0.95	0.95	22
	IGNORE	1.00	0.87	0.93	23
	ACK	1.00	1.00	1.00	3
	CHT	0.83	0.91	0.87	11
	DP	0.00	0.00	0.00	1
	accuracy			0.93	205
	macro avg.	0.81	0.82	0.81	205
	weighted avg.	0.94	0.93	0.93	205

Table 5: Detailed Annotation Report for **Polish Second Group** Annotators

Annotator	**Classes**	**Precision**	**Recall**	**F1-score**	**Frequency**
Second Group First Annotator	DA	0.93	0.97	0.95	339
	IND	0.85	0.76	0.80	45
	CR	0.82	0.45	0.58	20
	DPR	0.87	0.81	0.84	32
	IGNORE	0.82	0.56	0.67	25
	ACK	1.00	0.93	0.97	15
	CHT	0.28	0.50	0.36	10
	DP	0.40	1.00	0.57	2
	MOTIV	1.00	1.00	1.00	1
	accuracy			0.89	489
	macro avg.	0.77	0.78	0.75	489
	weighted avg.	0.89	0.89	0.88	489
Second Group Second Annotator	DA	0.99	0.78	0.88	339
	IND	0.49	0.84	0.62	45
	CR	0.92	0.55	0.69	20
	DPR	0.68	0.84	0.75	32
	IGNORE	0.43	0.92	0.58	25
	ACK	0.88	1.00	0.94	15
	CHT	0.28	0.50	0.36	10
	DP	1.00	1.00	1.00	2
	MOTIV	1.00	1.00	1.00	1
	accuracy			0.79	489
	macro avg.	0.74	0.83	0.76	489
	weighted avg.	0.88	0.79	0.81	489
Main Annotator	DA	0.92	0.95	0.94	339
	IND	0.83	0.64	0.73	45
	CR	0.69	1.00	0.82	20
	DPR	0.85	0.88	0.86	32
	IGNORE	0.71	0.68	0.69	25
	ACK	0.86	0.40	0.55	15
	CHT	0.88	0.70	0.78	10
	DP	0.50	0.50	0.50	2
	MOTIV	1.00	1.00	1.00	1
	accuracy			0.88	489
	macro avg.	0.80	0.75	0.76	489
	weighted avg.	0.88	0.88	0.88	489

The previous inter-annotator reliability study was carried out on the full taxonomy of question responses. However, we also performed inter-annotator reliability tests on several subsets of the taxonomy, to learn which subsets of the taxonomy can be reliably annotated. We also used Cohen's Kappa score for this task.

English The detailed Cohen's Kappa scores on different subsets of the taxonomy for English are presented in Table 6. As shown in the table, the response types DA, CR, ACK, DPR, DP, MOTIV were annotated with almost perfect agreement level (above 0.9) (McHugh, 2012) between annotators in both groups of the experiment. However, the response types such as IGNORE, CHT, IND caused a sharp decrease in the agreement level. The indirect answer (IND) is the one that drops the agreement level between annotators significantly.

Table 6: English inter-annotator reliability on subsets of the taxonomy, Cohen's Kappa score

Subset of Taxonomy	**1st vs. 2nd**	**1st vs. 3rd**
DA, CR	0.9816	1.0
DA, CR, ACK	0.9710	1.0
DA, CR, ACK, DPR	0.9681	0.9489
DA, CR, ACK, DPR, DP	0.9686	0.9489
DA, CR, ACK, DPR, DP, MOTIV	0.9692	0.9489
DA, CR, ACK, DPR, DP, MOTIV, IGNORE	0.8973	0.8755
DA, CR, ACK, DPR, DP, MOTIV, IGNORE, CHT	0.8739	0.8391
DA, CR, ACK, DPR, DP, MOTIV, IGNORE, CHT, IND	0.7183	0.6358
DA, CR, ACK, DPR, DP, MOTIV, IGNORE, CHT, IND, OTHER	0.7052	0.6430

Polish The agreement level among annotators on different subsets of the taxonomy for two groups of Polish annotation are displayed in Table 7 and Table 8 respectively. Comparing the overall results on two tables, we found that the agreement level among the annotators in the first group is generally higher than that of the second group. In the first group, the response types DA, CR, ACK, DPR, DP, MOTIV were annotated with a strong agreement level (0.8–0.9) (McHugh, 2012) between first and the main annotators, and also between the second and the main annotators. However, those response types were annotated with a moderate agreement level (0.60–0.79) (McHugh, 2012) between the first and the second annotators. As for the second group in Table 8, the response types DA, CR, ACK, DPR, DP, MOTIV were annotated with a moderate agreement level (0.60–0.79) nearly among all annotators. In both groups, the agreement level dropped evidently when IGNORE, CHT, IND were added.

To sum up, response types such as DA, CR, ACK, DPR, DP, MOTIV can be reliably annotated by all annotators in both languages, whereas the response types such as IGNORE, CHT, IND cause more confusion to the annotators. Among all response types, the indirect answer (IND) is the one that is most difficult to annotate.

Disagreement analysis

For English:

Among the commonly annotated 607 BNC Q-Rs, there are 108 cases where annotation disagreements between two annotators occurred as shown in Table 9. The main disagreements concerned DA versus IND (52), IGNORE versus CHT/ACK/DP/DA/DPR/IND (33), and ACK versus

Table 7: Polish First Group inter-annotator reliability on subsets of the taxonomy, Cohen's Kappa score

Subset of Taxonomy	**1st vs. 2nd**	**1st vs. main**	**2nd vs. main**
DA, CR	0.7882	0.8214	0.8074
DA, CR, ACK	0.7882	0.8214	0.8010
DA, CR, ACK, DPR	0.7855	0.8343	0.8781
DA, CR, ACK, DPR, DP	0.7582	0.8238	0.8449
DA, CR, ACK, DPR, DP, MOTIV	0.7582	0.8238	0.8449
DA, CR, ACK, DPR, DP, MOTIV, IGNORE	0.6867	0.7515	0.8498
DA, CR, ACK, DPR, DP, MOTIV, IGNORE, CHT	0.6360	0.6957	0.7863
DA, CR, ACK, DPR, DP, MOTIV, IGNORE, CHT, IND	0.4810	0.5315	0.6662
DA, CR, ACK, DPR, DP, MOTIV, IGNORE, CHT, IND, OTHER	0.4588	0.5121	0.6579

Table 8: Polish Second Group inter-annotator reliability on subsets of the taxonomy, Cohen's Kappa score

Subset of Taxonomy	**1st vs. 2nd**	**1st vs. main**	**2nd vs. main**
DA, CR	0.7525	0.6694	0.6522
DA, CR, ACK	0.8351	0.5901	0.5779
DA, CR, ACK, DPR	0.7652	0.6651	0.6399
DA, CR, ACK, DPR, DP	0.7612	0.6661	0.6406
DA, CR, ACK, DPR, DP, MOTIV	0.7648	0.6712	0.6462
DA, CR, ACK, DPR, DP, MOTIV, IGNORE	0.7040	0.6047	0.6429
DA, CR, ACK, DPR, DP, MOTIV, IGNORE, CHT	0.6220	0.5604	0.6025
DA, CR, ACK, DPR, DP, MOTIV, IGNORE, CHT, IND	0.5414	0.4738	0.5467

OTHER/DA/DPR/CHT (5), as exemplified in (30). Invariably, the direct/indirect disagreements occurred with 'why', 'how' and 'what is X doing' questions, where answers are by and large sentential and for which there has been significant controversy in the theoretical literature on how to characterize answerhood (Kuipers and Wiśniewski, 1994; Asher and Lascarides, 1998).

Table 9: Disagreement cases for English

Disagreement types	**Frequency**	**Disagreement types**	**Frequency**
DA-IND	**52**	DA-CR	1
DA-IGNORE	8	IGNORE-CHT	7
DA-DPR	4	IGNORE-DPR	3
DA-CHT	2	IGNORE-ACK	2
IND-CR	3	IGNORE-DP	3
IND-DP	2	ACK-DA	1
IND-DPR	3	ACK-CHT	1
IND-IGNORE	9	ACK-OTHER	3
IND-CHT	2	CR-OTHER	2
		SUM	**108**

(30) a. ANON 1: When did the bus service start to <unclear> then?

MANSIE FLAWS: **Oh it was a while after we started.** [*DA vs. IND, resolved to IND*]

b. ANN: That's not very nice.

STUART: It is.

ANN: No It isn't.

STUART: Well it is. Why isn't it?

ANN: **Cos it isn't.** [*DA v. IGNORE, resolved to DA.*]

c. JOHN: So lock erm how would you spell sock?

SIMON: **<laugh> smelly er smelly** [*IGNORE v. CHT, resolved to IGNORE.*]

JOHN: How would you smell sock then?

d. JOHN: Can you spell box?

SIMON: **Mhm.** [*ACK v. OTHER, resolved to DA, after consideration of surrounding context.*]

In the above conversations, (30a) is an example of DA versus IND, where the first annotator categorized it as IND, while the second person annotated it as DA. After discussion, we decided to classify it as IND given that a certain amount of inference is needed to know the exact time of the bus service. For (30b), the first annotator annotated it as IGNORE, while the second annotator marked it as DA, however, after discussion, we decided that it should be categorized in DA since the response emphasizes the fact that "because it is actually not nice ". For (30c), the first annotator annotated the answer as IGNORE, while the second person categorized it as CHT, and after discussion, we keep IGNORE as the correct annotation since the answer is also related to the main topic "sock". (30d) is an example of ACK versus OTHER, where the first annotator annotated it as OTHER, while the second annotator treated it as ACK. However, as a result of considering the surrounding context, we concluded that it is actually a direct answer to the question.

For Polish: For the whole annotated sample, we observed 41 cases with disagreement between all three annotators (as shown in Table 10). The main disagreements concerned DA versus DPR (12), which is a notable difference by comparison with the English data.[15] We also observed some DA versus IND disagreements but much less common (4). It is also the case that the IGNORE category appears often in the disagreements summary (versus DA, CR, IND, CHT, and DPR).

Among the analyzed disagreement cases, two are especially interesting as the disagreement of all annotators is observed for consecutive turns in a dialogue. The first problematic case is for [016O, 62–65]. A and B are discussing B's application for a scholarship.

(31) A: a w tej twojej szkole ty jako <PAUSE> twoja kandydatura została złożona tylko <PAUSE> czy jeszcze jakiś innych osób też [*and in your school it is you <PAUSE> you are the only candidate <PAUSE> or maybe there are some other people who also applied*]

[15] We hypothesize that the reason for this may be the background of annotators as logicians. From a logical perspective the exhaustiveness of an answer is important (see e.g. Wiśniewski, 2013). Thus, certain partial answers provided by dialogue participants were probably tagged as DPR. This may be due to the fact that partial answers were not explicitly pointed out in the guidelines.

Table 10: Disagreement cases for Polish (without the main annotator)

Disagreement types	**Frequency**	**Disagreement types**	**Frequency**
ACK-CHT	1	CHT-IND	2
DA-IND	4	IGNORE-DA	1
DA-DPR	**12**	IGNORE-CR	1
DA-CR	3	IGNORE-CHT	1
DA-CHT	3	IGNORE-IND	1
CR-DP	1	IGNORE-DPR	2
CR-CHT	1	IND-DPR	1
CR-DPR	1	DP-CR	1
CR-IND	1	OTHER-CHT	1
CR-CHT	3	**SUM**	**41**

B: co ty <PAUSE> nie no nie dowiadywałam się <PAUSE> wiem że z mojej grupy tylko ja jestem [*oh stop <PAUSE> I didn't check <PAUSE> I know only that from my group it was only me*]

A: masz konkurencję [*so you have some competition*]

B: yyy <PAUSE> z całej tej <PAUSE> szkoły ? nie wiem <PAUSE> na przykład od Marty <PAUSE> mogłabym się Marty zapytać właściwie <PAUSE> bo od Mar <PAUSE> Marta nie chciała jechać <PAUSE> właściwie to nie wiem dlaczego <PAUSE> ale już aż mi było <PAUSE> ty ja ją tak namawiałam <PAUSE> tak ją prosiłam <PAUSE> potem <PAUSE> ona i tak <PAUSE> tak wiesz to zlała <PAUSE> nie wiem dlaczego nie chciała pojechać [*yyy <PAUSE> from the whole <PAUSE> school? I don't know <PAUSE> for example Martha <PAUSE> actually I could ask Martha <PAUSE> because from Mar <PAUSE> Martha didn't want to go <PAUSE> actually I do not know why <PAUSE> but for me it was <PAUSE> you know I have tried to convince her <PAUSE> I have asked her <PAUSE> and then <PAUSE> she after all <PAUSE> you know, she just ignored it <PAUSE> I do not know why she didn't want to go*]

In this case, the disagreement between annotators was whether the first B's utterance should be classified as 'it is difficult to provide an answer' (DPR) or as an indirect answer (IND). As for the second B's utterance, the suggested types were DPR and IGNORE.

Another example where the disagreement was observed for two consecutive utterances is [01AO, 256–259]. Most probably, this is caused by the fact that four participants took part in this dialogue (which makes an interpretation of question responses much more difficult).

(32) B: ciekawe ile kasy dają [*I am wondering how much money they can give you*]

C: ciekawe ile kasy dają [*I am wondering how much money they can give you*]

A: no dawają ci tyle co na tym na [*well they give you the same that in that*]

D: w sklepie w kerfurze że po siedem złotych mówiła [*she said that in this shop in kerfur it is seven*]

Here C's utterance was tagged as OTHER, DP, and CR. It seems that in this case, C's utterance may be treated as a simple repetition of B's question, and as such, it should not be recognized as DP. As for A's utterance, it was tagged as IND, DPR, and DA by the annotators. In this case, the answer does not require any form of inference. It simply states that it will be the same amount of money you can earn in certain places. The place and the amount of money are then pointed out by the following D's utterance. That speaks for interpreting A's utterance as a DA (however, a partial one).[16]

8. Formal Analysis

There is a two-way relationship between corpus studies of questions and responses and formal semantic theories of questions and of dialogue. Notions from the latter play an important role in the design of the former. And one can strive to show that the categories posited are coherent formally using formal theories. Conversely, the ability to fully describe the data that emerges from such corpus studies can be used as a means for evaluating different approaches. Our aim in this section is to address both directions alluded to above.

Our explication is formulated using the frameworks of TTR (Cooper and Ginzburg, 2015; Cooper, 2023) (for the semantic ontology) and KoS (Ginzburg, 2012; Ginzburg et al., 2020) (for the theory of dialogue context); the relevant notions of TTR are sketched in Appendix A, whereas those of KoS are introduced in the text.

8.1 The classes DA, DP, IND

We assume that questions are propositional abstracts—extensive motivation for this view is provided in (Ginzburg, 1995; Ginzburg and Sag, 2000; Krifka, 2001); the particular implementation of this view in TTR can be found in (Ginzburg, 2012; Cooper and Ginzburg, 2015).[17]

(33) exemplifies the denotations (contents) we can assign to a unary, binary *wh*-interrogative and to polar questions. We use r_{ds} here to represent the record that models the described situation in the context. The meaning of the interrogative would be a function defined on contexts which provide the described situation and which return as contents the functions given in (33). The unary question ranges over instantiations by persons of the proposition "x runs in situation r_{ds}". The binary question ranges over pairs of persons x and things y that instantiate the proposition "x touches y in situation r_{ds}":

(33) a. who ran $\mapsto$

$$\lambda r{:}\begin{bmatrix}\text{x:Ind}\\ \text{rest:person(x)}\end{bmatrix}\left(\begin{bmatrix}\text{sit} & = & r_{ds}\\ \text{sit-type} & = & \left[\text{c:run}(r.\text{x})\right]\end{bmatrix}\right)$$

b. who touched what $\mapsto$

[16] As suggested by an anonymous reviewer for *Dialogue and Discourse*, it is plausible that in the discussed case A's intention was to provide a complete answer but this was interrupted by D.

[17] A variant on this view motivated by data from Boolean connectives and adjectives can be found in (Ginzburg et al., 2014).

$$\lambda r{:}\begin{bmatrix}\text{x:Ind}\\ \text{rest1:person(x)}\\ \text{y:Ind}\\ \text{rest2:thing(y)}\end{bmatrix}\left(\begin{bmatrix}\text{sit} & = & r_{ds}\\ \text{sit-type} & = & \left[\text{c:touch}(r.\text{x},r.\text{y})\right]\end{bmatrix}\right)$$

c. Did Bo run $\mapsto$

$$\lambda r{:}\mathit{Rec}\left(\begin{bmatrix}\text{sit} & = & r_{ds}\\ \text{sit-type} & = & [\ \text{c} : \text{run(bo)}\ ]\end{bmatrix}\right)$$

d. Didn't Bo run $\mapsto$

$$\lambda r{:}\mathit{Rec}\left(\begin{bmatrix}\text{sit} & = & r_{ds}\\ \text{sit-type} & = & [\ \text{c} : \neg\text{run(bo)}\ ]\end{bmatrix}\right)$$

Polar questions are analyzed, following an initial proposal of Ginzburg and Sag (2000), as 0-ary abstracts, which in TTR is a question whose domain is the empty record type [] (that is, the type *Rec* of records).[18] This makes a 0-ary abstract a constant function from the universe of all records. It allows to distinguish the denotations of positive and negative polar questions, as exemplified in (33c,d) and as motivated by a variety of linguistic phenomena (Hoepelmann, 1983; Cooper and Ginzburg, 2012). At the same time, it ensures that the answerhood relations they give rise to are (truth conditionally) equivalent, given that the *simple answerhood* relations they give rise to are equivalent and other answerhood relations are defined in terms of these.[19] Simple answerhood is the range of the propositional abstract, plus their negations. We exemplify what this amounts to for some cases in (34), using as we do mostly in the sequel familiar λ-notation for wh-questions and p?-notation for polar questions, rather than the official TTR notation above:[20]

(34) a. AtomAns($p?$) = $\{p\}$

b. AtomAns($\neg p?$) = $\{\neg p\}$

c. AtomAns($\lambda x.P(x)$) = $\{P(a), P(b), \ldots,\}$

d. NegAtomAns(q) = $\{p|\exists p_1 \in AtomAns(q), p = \neg p_1\}$

e. SimpleAns(q) = AtomAns(q) $\cup$ NegAtomAns(q)

Assuming questions to be propositional abstracts means that they can be used to *underspecify* answerhood. This is important given that NL requires a variety of answerhood notions, both for classifying responses and also for the role questions play as arguments to predicates such as 'know', 'tell', and 'depends', which in turn play a role in associated discourse reasoning (Groenendijk and

[18] This is the type all records satisfy, since it places no constraints on them.

[19] The need for such truth conditional equivalence is motivated *inter alia* by inferences such as the following:

(i) Jill knows whether Bo left.

(ii) Hence, Jill knows whether Bo did not leave.

[20] As Cooper and Ginzburg, 2015, §7.1 explain, the equivalence between the simple answerhoods of positive and negative polar answers follows because the negation operator on types $\neg$ satisfies for any s, T that— $s : T$ iff $s : \neg\neg T$, though T and $\neg\neg T$ are distinct types.

Stokhof, 1997; Wiśniewski, 2015). In fact, *simple answerhood*, though it has good coverage in practice, is not sufficient. It does not accommodate conditional, weakly modalized, and quantificational answers, all of which are pervasive in actual linguistic use (Ginzburg and Sag, 2000):

(35) a. Christopher: Can I have some ice-cream then?
Dorothy: you can do if there is any. (BNC)

b. Anon: Are you voting for Tory?
Denise: I might. (BNC, slightly modified)

c. How many players are getting these kind of opportunities to develop their potential? Not many. (The Guardian, Nov 2, 2018)

d. Dorothy: What did grandma have to catch?
Christopher: A bus. (BNC, slightly modified)

e. Elinor: Where are you going to hide it?
Tim: Somewhere you can't have it.

Thus, we suggest that the semantic notion relevant to direct answerhood is the relation *aboutness*—a relation between propositions and questions that any speaker of a given language can recognize, independently of domain knowledge and of the goals underlying an interaction.

The most detailed discussion of Aboutness we are aware of is (Ginzburg and Sag, 2000, pp. 129–149), which offers (36a) (reformulated here in TTR)[21] as a characterization of Aboutness that can accommodate data such as (35).[22] This requires the situational type component of the proposition to be a subtype of the join of the situational type of the question's simple answer set. As it stands, this definition allows in principle very informationally strong types as direct answers, since nothing bounds the proposition from above. Plausible upper bounds for direct answerhood familiar in the semantics of questions from the classic proposal of Karttunen (1977) are the meets of the question's atomic and negative atomic answer set.[23] This condition is formulated in (36b):[24,25]

[21] See Appendix A for some additional details.

[22] Ginzburg (1995) suggested that Aboutness is closed under conditionalization: i.e., for any r, p if p is about q, then so is if $r,\ then\ p$:

(i) A: Who will win tomorrow's match? B: If it isn't raining, the French.

(ii) A: Did someone switch the oven off? B: Unless you explicitly told them to, no one did.

The definition given in the text covers non–conditionalized answers. One crude strategy to obtain the latter, as proposed by Ginzburg (1995), is to extend the definition for non-conditionalized answers by closing it under conditionalization.

[23] For a polar question p? the meets of the question's atomic and negative atomic answer set are respectively p and $\neg p$, whereas for a wh–question $\lambda x.P(x)$ (e.g., 'who left') they are respectively $\bigwedge P(a_i)$ ('Bo left and Millie left ...'), whereas $\bigwedge \neg P(a_i)$ ('Bo did not leave and Millie did not leave ..., i.e., equivalent to 'No one left').

[24] For a wide ranging discussion of a variety of answerhood relations, see (Wiśniewski, 2015). He leaves the composition of his "base answer set", the *Principal possible answers* (PPAs), as a parameter of the theory, to be fixed independently from the questions, since his account is stated in an artificial logical language that is not directly tied to linguistic forms. Hence, his account is compatible in principle with most semantic approaches to questions.

[25] Our use of subtyping as a means of characterizing aboutness reflects that, as an anonymous reviewer for *Dialogue and Discourse* points out, both direct and indirect answerhood involve inference. As we discuss below, for the latter the notion of inference is an agent-relative notion.

(36) For $p = \begin{bmatrix}\text{sit} = \text{s}_1 \\ \text{sit-type} = \text{T}_1\end{bmatrix}$: Prop, q = (r : T_2) $\begin{bmatrix}\text{sit} = \text{s}_1 \\ \text{sit-type} = \text{T}_3\end{bmatrix}$: Question,

a. $About(p, q)$ holds iff $T_1 \sqsubseteq \bigvee\{T|\exists p'[p' : Prop \wedge SimpleAns(p', q) \wedge T = p'.\text{sit-type}]\}$

b. $DirectAns(p, q)$ holds iff $About(p, q)$ and either

(i) $\bigwedge AtomAns(q) \sqsubseteq T_1$

or

(ii) $\bigwedge NegAtomAns(q) \sqsubseteq T_1$

Despite the proposals mentioned above for explicating direct answerhood, a comprehensive, empirically-based, experimentally tested account for a variety of wh–words is still elusive and an important task for future work.

An additional important notion a theory of questions needs to provide for is a notion of *exhaustiveness* or *resolvedness*, though this is in general pragmatically parametrized (Ginzburg, 1995; Asher and Lascarides, 1998; van Rooy, 2003). Whether a response is resolving (or merely *goal fulfilling* without so doing) can determine whether the response will be accepted as sufficient to end discussion of the question or requires a follow up. Hence, the need for a finer–grained subdivision of the answer categories, as we hinted in footnote 7.

Given a notion of aboutness and some notion of (partial) exhaustiveness/resolvedness, one can then define question dependence (needed for the class DP), for instance, as in (37), though various alternative definitions have been proposed (Groenendijk and Stokhof, 1997; Groenendijk and Roelofsen, 2011; Wiśniewski, 2013). For all these definitions, as with aboutness, their coverage awaits testing on empirical data:

(37) $Depend(q_1, q_2)$ iff any proposition p such that p resolves $q2$, also satisfies p entails r for any r such that r is about $q1$, (Ginzburg, 2012, (61b), p. 57).

We have introduced answerhood notions corresponding to direct answerhood and to question–dependence, two of the three response categories we identified as *Question-Specific* in Section 3. Before we introduce the third notion, indirect answerhood, we sketch an account of dialogue context, which will allow us to integrate all three in a semantics for dialogue.

The simplest model of context, going back to Montague (1974), is one which specifies the existence of a speaker, addressing an addressee at a particular time. This can be captured in terms of the type in (38):

(38) $\begin{bmatrix}\text{spkr} & : & \textit{Ind} \\ \text{addr} & : & \textit{Ind} \\ \text{u-time} & : & \textit{Time} \\ \text{c}_{\text{utt}} & : & \text{addr(spkr,addr,u-time)}\end{bmatrix}$

However, over the last four decades it has become clearer how much more pervasive reference to context in interaction is. Expectations due to illocutionary acts—one act (querying, assertion, greeting) giving rise to anticipation of an appropriate response (answer, acceptance, counter–greeting),

also known as adjacency pairs (Schegloff, 2007). Extended interaction gives rise to shared assumptions or *presuppositions* (Stalnaker, 1978), whereas epistemic differences that remain to be resolved across participants—*questions under discussion* are a key notion in explaining coherence and various anaphoric processes (Ginzburg, 1994, 2012; Roberts, 1996). These considerations among several additional significant ones we discuss below lead work in KoS to two strategic moves: (i) instead of assuming a single context to be operative, a distributed notion is emergent from individual *Total Cognitive States* (TCS), one per participant. A TCS has two partitions, namely a *private*—about which we will not elaborate here—for details see (Larsson, 2002), and a *public* one.

(39) $$\text{TCS} = \begin{bmatrix} \text{public} & : & \mathit{DGBType} \\ \text{private} & : & \mathit{Private} \end{bmatrix}$$

(ii) we posit a significantly richer structure to represent each participant's view of publicized context, dubbed the *dialogue gameboard* (DGB), whose basic make up to process question–specific moves is given in (40):

(40) $$\mathit{DGBType} = \begin{bmatrix} \text{spkr} & : & \mathit{Ind} \\ \text{addr} & : & \mathit{Ind} \\ \text{utt-time} & : & \mathit{Time} \\ \text{c-utt} & : & \text{addressing(spkr,addr,utt-time)} \\ \text{facts} & : & \mathit{Set}(\mathit{Prop}) \\ \text{vis-sit} & = & \left[\text{foa} \;:\; \mathit{Ind} \vee \mathit{Rec}\right] : \mathit{RecType} \\ \text{moves} & : & \mathit{List}(\mathit{IllocProp}) \\ \text{qud} & : & \mathit{poset}(\mathit{Question}) \end{bmatrix}$$

Here *facts* represents the shared assumptions of the interlocutors—identified with a set of propositions. The parameters *spkr* and *addr* together with the addressing condition (at a given time) track verbal turns and mutual engagement. The remaining fields concern locutionary and illocutionary interaction. Within *moves* the first element has a special status given its use to capture adjacency pair coherence and it is referred to as *LatestMove*. The current question under discussion is tracked in the *qud* field, whose data type is a partially ordered set (*poset*). *Vis-sit* represents the visual situation of an agent, including his or her focus of attention (*foa*), which can be an object (*Ind*), or a situation or event (*Sit*), relevant *inter alia* for processing gestural answers.

We call a mapping between DGB types a *conversational rule*—Conversational rules are the means for specifying how DGBs evolve. The types specifying its domain and its range we dub, respectively, the *pre(conditions)* and the *effects*, both of which are subtypes of DGBType: they apply to a subclass of records that constitute possible DGBs and modify them to records that constitute possible DGBs. Conversational rules are written here in a form where the preconditions represent information specific to the preconditions of this particular interaction type and the effects represent those aspects of the preconditions that have changed.

The first conversational rule we formulate relates to the basic effect a query has on the DGB—as a consequence of a query a question becomes the maximal element of QUD:

(41) Ask QUD-incrementation: given a question q and Ask(A, B, q) being the LatestMove, one can update QUD with q as MaxQUD.

$$\left[\begin{array}{lcl} \text{pre} & : & \left[\begin{array}{lcl} \text{q} & : & \text{Question} \\ \text{LatestMove = Ask(spkr, addr, q)} & : & \text{LocProp} \end{array}\right] \\ \text{effects} & : & \left[\text{QUD} = \left\langle \text{q, pre.QUD} \right\rangle \ : \ \text{poset(Question)}\right] \end{array}\right]$$

With this initial view of context and context change in hand, we can return to discuss indirect answerhood. The notion of *direct* answer is clearly complex and, as we have indicated, probably needs, at least for dialogue management purposes, to be refined. With indirect answers the situation seems even more tricky, which in part reflects why this category is one of those with most inter-annotator variability. Indirectness encapsulates various notions, as we have already discussed in Section 2. There is a considerable literature on indirect speech acts, building on and reacting to initial notions from Grice (1975) and Searle (1975). Roughly speaking, these involve cases where the speaker's intention is not transparently reflected in an utterance's *grammatically governed content*—the content whose resolution is driven by conventional mechanisms.[26] The classic Gricean model involves *initial* recognition of a literal content (corresponding to what we have referred to above as 'grammatically governed content')[27] and then, via domain–specific means, inference of the speaker's intention. Significant doubts about this time course, about the necessity of actually consulting a/the literal content, and what should be viewed as the literal/direct content have been debated extensively in the pragmatics literature, much of it in recent years on an experimental basis—for detailed review see (Noveck, 2018). Indirect speech acts are of course also an important theme in the AI planning literature, e.g., (Cohen and Perrault, 1979), incorporated in dialogue semantic frameworks in (Larsson, 2002; Asher and Lascarides, 2003; Ginzburg, 2012).

While a detailed analysis is beyond our scope here, one can distinguish at least two cases, which we might label as *shallow* and *deep* indirect answers. The former corresponds to cases like (11) and (13) repeated here as (42a,b) respectively, where the entailment of a direct answer is due to shallow shared knowledge (for (42a): find(a,b,t_1) $\rightarrow$ look_for(a,b,t_0), so by contraposition $\neg\exists$t look_for(a,b,t) $\rightarrow \neg$ find(a,b,t_1)) or to domain–independent erotetic reasoning (Wiśniewski, 2013), which adjusts the question asked to a close variant (Larsson, 1998) (e.g., $?\exists x.P(x) \rightarrow \lambda x.P(x)$, for (42b)). Some initial refinement of IND along these lines is hinted in footnote 8 above. This contrasts with the *deep* indirect answers, exemplified in (42c), which involve reasoning about the speaker's intentions, most often though not invariably based on domain–specific information. For detailed discussion of *deep* indirect answers within SDRT, see (Asher and Lascarides, 2001, 2003); for an account within KoS, see (Ginzburg, 2012, §8.3).

(42) a. Q: And also did you find my blue and green striped tie?

R: I haven't looked for it.

[26] By this we mean content whose contextual parameters are conventionally specified, e.g., 'Jill left' conventionally specifies predication of some concept of leaving applying to a person the speaker refers to as 'Jill'; resolving which concept of leaving and which Jill is less clearly rule–driven, though is a complicated mix of speaker/audience interaction, contextual salience etc.

[27] We use the latter somewhat pedantic term to differentiate it from the former, which has a variety of problematic associations. As will become clear in section 8.2, we do not assume that in general speaker and addressee need identically resolve even the grammatically governed content.

b. Q: Isn't your country seat there somewhere?

R: [Yes/No].
Stoke d'Abernon.

c. (Context: in queue for toilet on an aircraft)

ANON WOMAN: How desperate are you?

ME: (shrugs), Go ahead. (Ginzburg, 2012, p. 304)

Two basic conditions seem to characterize these cases: first, the indirect answer p is NOT a direct answer to the question q in the sense of the definition in (36b); second, p together with some shared knowledge, i.e., an element of FACTS for some dialogue gameboard dgb, the *bridging proposition* $bridgeprop$, entails r, which is a direct answer to q:[28]

(43) Given $p : Prop, q : Question, dgb : DGBType$ InDirectAns(p,q,dgb) iff ¬DirectAns(p,q) and there exist $bridgeprop, r : Prop$ such that DirectAns(r,q) and In(dgb.FACTS,bridgeprop) and $\rightarrow (p \wedge bridgeprop, r)$.

To exemplify: for (44a) asked by A who B knows needs to get up after sunrise, we could assume that the indirect answer p conjoined with (presumably shared) $bridgeprop$ entails r:[29]

(44) a. A: Is it time to rise? B: It is still dark outside.

b. $p = Dark(here, now)$

c. $bridgeprop$ = If it is dark here now, the time now is before A needs to rise.

d. $r = \neg NeedRise(A, now)$

We can now formulate a rule that explicate how answers and depended-upon questions get introduced in dialogue. This rule characterizes the contextual background of reactive queries and assertions—if q is MaxQUD, then subsequent to this *either* conversational participant may make a move which is either a (direct or indirect) answer or a question on which q depends).

(45) a. Given r : Question ∨ Prop, q : Question, dgb : DGBType, $QSpecific(r, q, dgb)$ iff DirectAns(r,q) ∨ IndirectAns(r,q,dgb) ∨ Depend(q,r)

[28] We leave open which notion of entailment, here denoted '→, is involved, whether directly relatable to the earlier subtyping or some other notion.

[29] An anonymous reviewer for *Dialogue and Discourse* asks whether p in such cases is the 'literal' content of the utterance or some strengthened version thereof such as some notion of speaker intended content, suggesting that in the latter case there might be significantly less need for indirect answerhood. Given that grammatically governed content is the input to repair processes (Ginzburg et al., 2003), it seems important for us to maintain p as a proposition that is explicitly not a direct answer if we wish to capture inter alia the clarificational potential from the addressee's perspective, as well as the speaker's choice in not explicitly uttering a direct answer. At the same time, we follow the reviewer's suggestion in offering a characterization that is not strictly at the propositional level, since it makes intrinsic use of shared knowledge in entailing the direct answer, whereas for direct answers we use information state–independent type subsumption. This and other questions by the reviewer concerning indirect answers have led to several reformulations of our earlier proposed characterizations of indirect answerhood.

b. QSPEC = $\left[\begin{array}{l} \text{pre} : \left[\text{QUD} = \left\langle \text{q, Q} \right\rangle : \text{poset(Question)}\right] \\ \text{effects} : \left[\begin{array}{l} \text{spkr} = \text{pre.spkr} \vee \text{pre.addr} : \text{Ind} \\ \text{addr} : \text{Ind} \\ \text{c}_{addr} : \neq(\text{addr,spkr}) \\ \text{p: Prop} \vee \text{Question} \\ \text{c1} : \text{QSpecific(p,q,pre)} \end{array}\right] \end{array}\right]$

8.2 The classes CR and ACK

MetaCommunicative utterances, including acknowledgements, clarification responses (CRs) (also known as *other repair* and as *other communication management*) and (metacommunicative) corrections are challenging for most existing frameworks for dialogue semantics. For a start, given the mismatch they reveal between the dialogue interlocutors, they require a *distributed* approach to context. This rules out accounts where all semantic rules are assumed to apply to the common ground, made prominent in the view of QUD due to Roberts (1996).[30] This was also the case for the view of discourse structure in earlier work in SDRT (e.g., Asher and Lascarides 1998, 2003). In more recent work (e.g., Lascarides and Asher 2009), SDRT adopts a view advocated in KoS and also in the framework of PTT (Poesio and Rieser, 2010) that associates a distinct contextual entity with each conversational participant.

A deeper challenge is that the analysis/generation of metacommunicative utterances requires access to the entire sign associated with a given interrogative utterance. This is for two main reasons. On the one hand, any constituent, certainly down to the word level can be the object of an acknowledgement and a clarification response, as exemplified for clarification responses in (46). Moreover, as discussed in detail in (Ginzburg, 2012), there are a variety of parallelism constraints relating to the form of such utterances that require reference to the non-semantic representation of the utterance. An illustration of this is given in (47) where the followup responses of two essentially synonymous questions turns out to be quite distinct:

(46) a. [George] Galloway [MP] is recorded reassuring his Excellency [Uday Hussein] that 'I'd like you to know we are with you 'til the end.' Who did he mean by 'we'? Who did he mean by 'you'? And what 'end' did he have in mind? He hasn't said. (From a report in the *Cambridge Varsity* by Jon Swaine, 17 February 2006)

b. Is The War Salvageable? That depends on what we mean by 'the war' and what we mean by 'salvage'. (Andrew Sullivan's Blog *The Daily Dish*, Sept, 2007)

(47) a. A: Do you fear him? B: Fear? (=What do you mean by 'fear' or Are you asking if I *fear* him) / #Afraid? / What do you mean 'afraid'?

b. A: Are you afraid of him? B: Afraid? (=What do you mean by "afraid"? or Are you asking if I am *afraid* of him) / #Fear?/What do you mean 'fear'?

[30] For a more refined stack–based discourse model, which distinguishes distinct participants' *commitments* see (Farkas and Bruce, 2010).

This issue, first discussed in some detail in (Ginzburg and Cooper, 2004), rules out the lion's share of logic–based frameworks where reasoning about coherence operates solely at the level of content. For instance, in SDRT the semantics/pragmatics interface has no access to linguistic form, but only to a partial description of the content that is derived from linguistic form. This has been argued to be necessary to ensure the decidability of SDRT's glue logic (see e.g., Asher and Lascarides 2003, p. 77).

In order to accommodate this class of utterances, it is crucial that the cognitive states keep track of the utterance associated with the question. In KoS this is handled via the field PENDING whose type (*LocProp*) is a record with two fields, one instantiated by an utterance token u, the other by an utterance type T_u (the sign classifying u); this allows *inter alia* access to the individual constituents of an utterance.

This leads to the following modified architecture for DGBs—they are distributed across dialogue participants (in other words—each participant is assigned their own DGB) and they include the field Pending consisting of ungrounded utterances:

(48) *DGBType* $\mapsto$

$$\begin{bmatrix} \text{spkr} & : & \textit{Ind} \\ \text{addr} & : & \textit{Ind} \\ \text{utt-time} & : & \textit{Time} \\ \text{c-utt} & : & \text{addressing(spkr,addr,utt-time)} \\ \text{facts} & : & \textit{Set}(\textit{Prop}) \\ \text{pending} & : & \textit{List}(\textit{LocProp}) \\ \text{moves} & : & \textit{List}(\textit{IllocProp}) \\ \text{qud} & : & \textit{poset}(\textit{Question}) \end{bmatrix}$$

Ginzburg and Cooper (2004); Purver (2004); Ginzburg (2012) show how to account for the main classes of CRs using rule schemas of the form "if u is the interrogative utterance and $u0$ is a constituent of u, allow responses that are *co-propositional*[31] with the clarification question $\text{CQ}^i(u0)$ into QUD.", where '$\text{CQ}^i(u0)$' is one of the three types of clarification question (repetition, confirmation, intended content) specified with respect to $u0$.

For instance, responses such as (46b) can be explicated in terms of the schema in (49):

(49) if A's utterance u is yet to be grounded and $u0$ is a sub-utterance of u, QUD can be updated with the question *What did A mean by u0*

More formally: the issue $q0$, *what did A mean by u0*, for a constituent $u0$ of the maximally pending utterance, A its speaker, can become the maximal element of QUD, licensing follow up utterances that are CoPropositional with $q0$. Assuming a propositional function view of questions, CoPropositionality allows in propositions from the range of $Range(q0)$ and questions whose range intersects $Range(q0)$. Since CoPropositionality is reflexive, this means in particular that the inferred clarification question is a possible follow up utterance, as are confirmations and corrections, as exemplified in (51).

[31] Here *CoPropositionality* for two questions means that, modulo their domain, the questions involve similar answers: for instance 'Whether Bo left', 'Who left', and 'Which student left' (assuming Bo is a student.) are all co-propositional.

(50) Parameter identification:

$$\begin{bmatrix} \text{pre} & : & \begin{bmatrix} \text{MaxPENDING} = \begin{bmatrix} \text{sit} = \text{u} \\ \text{sit-type} = \text{T}_u \end{bmatrix} : \text{LocProp} \\ \text{A} = \text{u.dgb-params.spkr} : \text{IND} \\ \text{u0} : \text{sign} \\ \text{c1} : \text{Member(u0,u.constits)} \end{bmatrix} \\ \text{effects} & : & \begin{bmatrix} \text{MaxQUD} = \lambda x \text{Mean(A,u0,x)} : \text{Question} \\ \text{LatestMove} : \text{LocProp} \\ \text{c1: CoPropositional(LatestMove.cont,MaxQUD)} \end{bmatrix} \end{bmatrix}$$

(51) a. $\lambda x.Mean(A, u0, x)$

b. ?Mean(A,u0,b) ('Did you mean Bo')

c. Mean(A,u0,c) ('You meant Chris')

8.3 The classes MOTIV, DPR, CHT, IGNORE

Łupkowski and Ginzburg (2016) suggest that common to all classes of evasion utterances is a lack of acceptance of $q1$ as an issue to be discussed. In MOTIV-type responses the need/desirability to discuss $q1$ is explicitly posed, in CHT-type responses there is an implicature that $q1$ is of lesser importance/urgency than r_2 (expressing either a proposition or a question), whereas for IGNORE type responses there is an implicature that q1 as such will not be addressed. Łupkowski and Ginzburg (2016) also note that whereas q_1 is not accepted for discussion, it remains implicitly in the context. In (52), where move (2) could involve either a MOTIV query (2a), or a CHT query (2b), the original question has definitely *not* been re-posed and yet B still has the option to address it, which s/he should be unable to do if it is not added to his/her context before (52(2)). Similar remarks *mutatis mutandis* apply to the DPR utterance in (52b):

(52) a. **A:** Who are you meeting next week?
B(2)**:** (2a) What's in it for you? / (2b) Who are *you* meeting next week?
A: I'm curious.
B: Aha.
A: Whatever.
B: Oh, OK, Jill.

b. A: When are you leaving? B: I don't know. A: Come on! B: Well, perhaps next week.

This basic characteristic can be captured in the cognitive state architecture discussed above, given that QUD is assumed to be partially ordered; this is a crucial difference from a view of QUD as a stack or similar (Roberts, 1996; Farkas and Bruce, 2010).

Concretely, Łupkowski and Ginzburg (2016) proposed to handle *metadiscursive* utterances such as MOTIV by viewing them as responses specific to the issue ?WishDiscuss(B,q) for a given question q and responder conversational participant B. This same approach can be applied to DPR, which Łupkowski and Ginzburg (2016) did not analyze, assuming that these involve responses specific to the issue λxKnowAnswer(x, q). We assume this formulation of the issue given the possiblity

of responses along these lines of 'Sam knows', 'You don't know?' etc.[32] In fact, we will deviate somewhat from the account of Łupkowski and Ginzburg (2016) in proposing a more uniform account than they did of all four classes for reasons we explain below. In order to do this, we will define a single type EvasiveResp that encompasses the commonalities between the four classes; each class will then be specified by merging EvasiveResp with information specific to that particular class. In all cases, in line with the fact that q remains accessible, as exemplified in (52), QUD is specified to include both q and a pertinent 'metaquestion'. An additional commonality for all except DPR is turn change, underspecified for QSPEC given that for the latter it is not required, whereas in these cases it is more or less essential for coherence; this specification will be defused for DPR by using asymmetric merge.

(53) EvasiveResp=
$$\left[\begin{array}{l} \text{pre} : \left[\text{QUD} = \left\langle \text{q1}, \text{Q} \right\rangle : \text{poset(Question)}\right] \\ \text{effects} : \left[\begin{array}{l} \text{spkr} = \text{pre.addr} : \text{Ind} \\ \text{addr} = \text{pre.spkr} : \text{Ind} \\ \text{r} : \text{Question} \vee \text{Prop} \\ \text{q2} : \text{Question} \\ \text{R: IllocRel} \\ \text{Moves} = \left\langle \text{R(spkr,addr,r)} \right\rangle \bigoplus \text{pre.Moves} : \text{list(LocProp)} \\ \text{c1} : \text{Qspecific(R(spkr,addr,r),q2)} \\ \text{QUD} = \left\langle \begin{array}{l} \text{Max} = \left\{ \text{q2,q1} \right\}, \\ \text{Q} \end{array} \right\rangle : \text{poset(Question)} \end{array}\right] \end{array}\right]$$

Given this, MOTIV and DPR are specified as follows:[33]

[32]Utterances like 'I don't know' and other DPR are differentiated from some other metadiscursive utterances in that the former can be used by the same speaker as a follow up, whereas the latter only if the speaker is correcting herself for having asked the question:

(i) A: Who should we invite?

(ii) . . . I don't know.

(iii) . . . # Do we need to talk about this now?

(iv) . . . # I don't wish to discuss this now.

Note also that 'I don't know' can be used as an editing phrase (Tian et al., 2015)—'She's I don't know 29.'.

[33]The basic idea of *merge* for record types is illustrated by the examples in (i,ii).

(i) $\left[\text{f}:T_1\right] \mathbin{\underset{\cdot}{\wedge}} \left[\text{g}:T_2\right] = \begin{bmatrix}\text{f}:T_1 \\ \text{g}:T_2\end{bmatrix}$

(ii) $\left[\text{f}:T_1\right] \mathbin{\underset{\cdot}{\wedge}} \left[\text{f}:T_2\right] = \left[\text{f}:T_1 \mathbin{\underset{\cdot}{\wedge}} T_2\right]$

In *asymmetric merge*, $T_1 \boxed{\mathbin{\underset{\cdot}{\wedge}}} T_2$, the second argument takes priority over the first, e.g.,

(iii) $\begin{bmatrix}\text{x}:T_1 \\ \text{y}:T_2\end{bmatrix} \boxed{\mathbin{\underset{\cdot}{\wedge}}} \left[\text{x}=a:T_1\right] = \begin{bmatrix}\text{x}=a:T_1 \\ \text{y}:T_2\end{bmatrix}$

(iv) $\begin{bmatrix}\text{x}=a:T_1 \\ \text{y}:T_2\end{bmatrix} \boxed{\mathbin{\underset{\cdot}{\wedge}}} \left[\text{x}=b:T_1\right] = \begin{bmatrix}\text{x}=b:T_1 \\ \text{y}:T_2\end{bmatrix}$

For a full definition which makes clear what the result is of merging any two arbitrary types, see (Cooper, 2012, 2023).

(54) a. MOTIV = EvasiveResp $\dot{\wedge}$ $\left[\text{effects} : \left[\text{q2} = ?\text{WishDiscuss(spkr,pre.MaxQUD)} : \text{Question}\right]\right]$

b. DPR = EvasiveResp $\boxed{\dot{\wedge}}$ $\left[\text{effects} : \begin{bmatrix}\text{spkr} = \text{pre.spkr} \vee \text{pre.addr} : \text{Ind} \\ \text{addr} : \text{Ind} \\ \text{c}_{addr} : \neq(\text{addr,spkr}) \\ \text{q2} = \lambda x\text{Know(x,pre.MaxQUD)} : \text{Question}\end{bmatrix}\right]$

With respect to both CHT and Ignore, we adopt a somewhat different perspective than that offered by Łupkowski and Ginzburg (2016), for both empirical and conceptual reasons. Considering the much larger dataset considered in this paper, their view of CHT seems too "cooperative" and that of IGNORE too "hostile". The analysis they offered for IGNORE built on an earlier analysis in (Ginzburg, 2012) intended to capture Gricean *irrelevance*, floutings of the Gricean maxim of relevance as in (55). That analysis was designed to explain how the initial utterance in effect gets expunged from the DGB.

(55) A: Rozzo just gave a terrible talk. B: It's really hot and unpleasant here.

However, IGNOREs often occur in quite cooperative environments such as the MapTask, where under time pressure the response is driven by the observed situation. Indeed, Table 9 indicates that IGNOREs were most frequently confused with answers (direct and indirect) and with CHTs; the former datum suggests, therefore, that IGNOREs are susceptible to be viewed as addressing something related to the question asked. On the other hand, as far as CHT goes, the analysis of Łupkowski and Ginzburg (2016) was, arguably, too "cooperative". Łupkowski and Ginzburg (2016) assume that r_2 is constrained to be unifiable with q_1 via a question q_3 (e.g., q_1 = what do you (B) like? r_2 = what do you (A) like? q_3 = Who likes what?). This assumption was motivated by a certain paralellism that seems to occur frequently between q_1 and r_2 when the latter has the form of a question. Imposing this condition, which requires a question inference mechanism for testing this unifiability, significantly constrains the CHT relation. However, in the more general case, where responses are not constrained to be questions, this condition seems less justified and, even focussing on question responses, the constructed example (56) seems quite natural:[34]

(56) A: When are you going to respond to the allegations? B: Anyway, when are we going to get credit for our world leading vaccination program?

The simplest analysis for IGNORE would make the pertinent meta-question be an arbitrary question about entities in the visual situation. Similarly, for CHT the simplest analysis would involve allowing a response specific to an arbitrary question. The obvious problem this would raise in both cases is massive ambiguity since many responses from other classes would be analyzable in such terms. To avoid this problem, we need to introduce an additional restriction, for instance along the lines of the afore-mentioned *irrelevance*; in other words, lack of coherence with the current context. What would this amount to? Being neither QSpecific with respect to q_1 uttered by A to B, nor being co-propositional with a clarification question generated by q_1's utterance, nor QSpecific with respect to ?WishDiscuss(B,q1) or λxKnowAnswer(x,q1). Putting these conditions together amounts to the IrRel relation of Ginzburg (2012), which holds between an utterance and a DGB.

[34]The example is constructed, but familiar to anyone following the British political scene in early 2022.

Given this, we formulate the rules for CHT and IGNORE as in (57a) and (57b). The fact that in both cases the topic addressed is irrelevant(IrRel) to the (precondition) DGB in the sense just discussed captures a similarity between the two. At the same time, there is also a significant difference in that IGNORE intrinsically uses material from the DGB, namely at least one entity from the visual situation as a constituent of the propositional nucleus of the question to establish coherence with the question posed. A further difference between the two–and deviation from (Łupkowski and Ginzburg, 2016)—is an emergent presupposition in the case of CHT that the responder does not wish to discuss q_1.

(57) a. CHT = EvasiveResp $\dot{\wedge}$
$$\begin{bmatrix} \text{effects} : \begin{bmatrix} \text{q2} : \text{Question} \\ \text{c3} : \text{IrRel(q2,pre)} \\ \text{Facts} = \text{pre.Facts} \cup \\ \left\{ \neg \text{ WishDiscuss(spkr,pre.MaxQUD)} \right\} \end{bmatrix} \end{bmatrix}$$

b. IGNORE = EvasiveResp $\dot{\wedge}$
$$\begin{bmatrix} \text{effects} : \begin{bmatrix} \text{a} : \text{Ind} \\ \text{c4} : \text{In(VisSit,a)} \\ \text{G1} : \text{Type} \\ \text{P} : \text{(Ind)RecType} \\ \text{q2} = \text{(G1)} \begin{bmatrix} \text{sit} = s \\ \text{sit-type} = \left[\text{c} : \text{P(a)} \right] \end{bmatrix} : \text{Question} \\ \text{c3} : \text{IrRel(q2,pre)} \end{bmatrix} \end{bmatrix}$$

9. Conclusions and Future Work

In this paper, we have presented an initial study for what is, as far as we are aware, the first, detailed, formally underpinned characterization of the response space of questions. Concretely, our initial hypothesis, stated in the introduction as (1) is repeated here as (58):

(58)(H) *Main hypothesis*: responses drawn from or concerning the LG query classes plus direct answerhood exhaust the response space of a query.

We think the data provided in previous sections validates this hypothesis, though we have made some small adjustments—conflating several classes. Achieving such a characterization is a fundamental challenge for semantics with a very wide variety of applications. It establishes theoretical benchmarks for theories of dialogue, for dialogue systems, and for semantic theories of questions.

Apart from the need to scale up the evidence quantitatively, we are currently engaged in work on the following strands:

- Extending the characterisation of response spaces to other moves: we have partitioned the response space into question–specific and non–question-specific (Metacomm, CHT, IGNORE, MOTIV, DPR). This suggests that other moves such as assertions and commands can be characterized in similar terms, where the non–question-specific class is applicable to all.

- The account we have developed is domain general, abstracting over differences between different conversational types/genres/language games etc. To what extent the current account will change once one takes such differences into account is an important question.

- Cross-question type comparison: the Q-R pairs annotated in the current study were selected randomly, whereas it is clearly of interest to consider the distribution of responses relative to fixed classes of questions (e.g., different classes of wh–questions, polar questions etc.)

- Apply machine learning to acquire the response classification scheme: Yusupujiang et al. (2022) provide an initial study comparing both classical machine learning algorithms as well as pretrained language models such as BERT (Devlin et al., 2018). This achieves encouraging results on some classes (e.g., DA and CR), while struggling with heavily inference-based classes like indirect answers, and IGNORE/CHT. This learnability trend is closely in line with that achieved by the human annotators in the current paper.

- Spoken dialogue system implementation: we plan to test the usability of these categories in dialogue systems. For this, one needs dialogue systems with sophisticated NLU, along the lines sketched in (Maraev et al., 2018, 2020).

- Cross-linguistic testing: a significant challenge is how to test the classification with languages lacking large or even hardly any speech corpora. We anticipate using online games with a purpose to this end (see e.g. Łupkowski and Ignaszak 2017; Łupkowski et al. 2018; Yusupujiang and Ginzburg 2020). For an initial study concerning the response space of queries in Uyghur, see (Yusupujiang and Ginzburg, 2022).

Finally, it is worth mentioning that at least part of our response typology can be be straightforwardly related to one of the well known annotation standards for dialogues, namely the ISO 24617-2 (Bunt, 2019).[35] The standard focuses on functional segments of dialogue acts. These segments are understood as "minimal stretches of communicative behavior that have a communicative function, 'minimal' in the sense of not including material that does not contribute to the expression of the function or the semantic content of the dialogue act" (Bunt, 2019, p. 4). When it comes to the general-purpose functions, dialogue acts may be information-providing (making certain information available to the addressee) or information-seeking (where information to be obtained can be of any kind, relating to the underlying task or activity, or even relating to the interaction itself). Among the information-providing functions, two sub-categories are distinguished: answer functions (where the speaker is providing information in response to an information need) and informing functions (where the speaker wants the addressee to know or be aware of something. One may notice that parts of our typology relate to the scope of the information-providing functions. DA and IND fall under answer functions, and ACK, IDK, DPR as well as CR may be categorized as informing functions. What would be interesting is to find a place for evasive responses in the DIT++ scheme (probably among the dimension-specific functions). What remains an open question is how to incorporate question-responses into the aforementioned scheme.

[35] Stemming from the DIT++ annotation scheme (Bunt, 2009).

10. Acknowledgements

This work is supported by a public grant overseen by the French National Research Agency (ANR) as part of the program "Investissements d'Avenir" (reference: ANR-10-LABX-0083). It contributes to the IdEx Université de Paris - ANR-18-IDEX-0001. We also acknowledge a senior fellowship from the Institut Universitaire de France to the first author, which funded the internships of Yusupujiang, Li, and Ren at LLF.

References

Anne H. Anderson, Miles Bader, Ellen Gurman Bard, Elizabeth H. Boyle, Gwyneth M. Doherty, Simon C. Garrod, Stephen D. Isard, Jacqueline C. Kowtko, Jan M. McAllister, Jim Miller, Catherine F. Sotillo, Henry S. Thompson, and Regina Weinert. The HCRC Map Task Corpus. *Language and Speech*, 34(4):351–366, 1991.

Aristotle. *Eudemian Ethics*. Cambridge University Press, 2012. Edited by Brad Inwood and Raphael Woolf.

Nicholas Asher and Alex Lascarides. Questions in dialogue. *Linguistics and Philosophy*, 21(3): 237–309, 1998.

Nicholas Asher and Alex Lascarides. Indirect speech acts. *Synthese*, 128(1):183–228, 2001.

Nicholas Asher and Alex Lascarides. *Logics of conversation*. Cambridge University Press, Cambridge, 2003.

John L. Austin. Truth. In James Urmson and Geoffrey J. Warnock, editors, *Philosophical Papers*. Oxford University Press, 1961. Paper originally published in 1950.

Jon Barwise and John Etchemendy. *The Liar*. Oxford University Press, New York, 1987.

Jon Barwise and John Perry. *Situations and Attitudes*. Bradford Books. MIT Press, Cambridge, 1983.

Ginger Berninger and Catherine Garvey. Relevant replies to questions: Answers versus evasions. *Journal of Psycholinguistic Research*, 10(4):403–420, 1981.

Harry Bunt. The DIT++ taxonomy for functional dialogue markup. In *AAMAS 2009 Workshop, Towards a Standard Markup Language for Embodied Dialogue Acts*, pages 13–24, 2009.

Harry Bunt. *Guidelines for using ISO standard 24617-2*. Tilburg University, Jan 2019. TiCC TR 2019–1.

Lou Burnard, editor. *Reference guide for the British National Corpus (XML Edition)*. Oxford University Computing Services on behalf of the BNC Consortium, 2007. URL `http://www.natcorp.ox.ac.uk/XMLedition/URG/`. acess 20.03.2017.

Jean Carletta. Assessing agreement on classification task: The kappa statistic. *Computational Linguistics*, 22(2):249–254, 1996.

Philip Cohen and Ray Perrault. Elements of a plan-based theory of speech acts. *Cognitive Science*, 3:177–212, 1979.

Robin Cooper. Type theory and semantics in flux. In Ruth Kempson, Nicholas Asher, and Tim Fernando, editors, *Handbook of the Philosophy of Science*, volume 14: Philosophy of Linguistics, pages 271–323. Elsevier, Amsterdam, 2012.

Robin Cooper. *From Perception to Communication: a Theory of Types for Action and Meaning*. Oxford University Press, 2023. URL `https://sites.google.com/site/typetheorywithrecords/drafts/`.

Robin Cooper and Jonathan Ginzburg. Negative inquisitiveness and alternatives-based negation. In Maria Aloni, Floris Roelofsen, Galit Weidman Sassoon, Katrin Schulz, Vadim Kimmelman, and Matthijs Westera, editors, *Proceedings of the 18th Amsterdam Colloquium*, 2012.

Robin Cooper and Jonathan Ginzburg. Type theory with records for natural language semantics. In Chris Fox and Shalom Lappin, editors, *Handbook of Contemporary Semantic Theory, second edition*, Oxford, 2015. Blackwell.

Cristian Danescu-Niculescu-Mizil and Lillian Lee. Chameleons in imagined conversations: A new approach to understanding coordination of linguistic style in dialogs. In *Proceedings of the 2nd workshop on cognitive modeling and computational linguistics*, pages 76–87. Association for Computational Linguistics, 2011.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

Nicholas J Enfield. Questions and responses in Lao. *Journal of Pragmatics*, 42(10):2649–2665, 2010.

Nicholas J Enfield, Tanya Stivers, Penelope Brown, Christina Englert, Katariina Harjunpää, Makoto Hayashi, Trine Heinemann, Gertie Hoymann, Tiina Keisanen, Mirka Rauniomaa, et al. Polar answers. *Journal of Linguistics*, 55(2):277–304, 2019.

Donka F Farkas and Kim B Bruce. On reacting to assertions and polar questions. *Journal of Semantics*, 27(1):81–118, 2010.

T. Fernando. Observing events and situations in time. *Linguistics and Philosophy*, 30(5):527–550, 2007.

Jonathan Ginzburg. An update semantics for dialogue. In H. Bunt, editor, *Proceedings of the 1st International Workshop on Computational Semantics*. ITK, Tilburg University, Tilburg, 1994.

Jonathan Ginzburg. Resolving questions, I. *Linguistics and Philosophy*, 18:459–527, 1995.

Jonathan Ginzburg. Situation semantics and the ontology of natural language. In Klaus von Heusinger, Claudia Maierborn, and Paul Portner, editors, *The Handbook of Semantics*. Walter de Gruyter, 2011.

Jonathan Ginzburg. *The Interactive Stance: Meaning for Conversation*. Oxford University Press, Oxford, 2012.

Jonathan Ginzburg and Robin Cooper. Clarification, ellipsis, and the nature of contextual updates. *Linguistics and Philosophy*, 27(3):297–366, 2004.

Jonathan Ginzburg and Ivan A. Sag. *Interrogative Investigations: the form, meaning and use of English Interrogatives*. Number 123 in CSLI Lecture Notes. CSLI Publications, Stanford: California, 2000.

Jonathan Ginzburg, Ivan Sag, and Matthew Purver. Integrating conversational move types in the grammar of conversation. *Perspectives on dialogue in the new millennium*, 114:25–42, 2003.

Jonathan Ginzburg, Robin Cooper, and Tim Fernando. Propositions, questions, and adjectives: a rich type theoretic approach. In *Proceedings of the EACL 2014 Workshop on Type Theory and Natural Language Semantics (TTNLS)*, pages 89–96, Gothenburg, Sweden, April 2014. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/W14-1411.

Jonathan Ginzburg, Chiara Mazzocconi, and Ye Tian. Laughter as language. *Glossa*, 5(1):104, 2020. doi: 10.5334/gjgl.1152.

Nancy Green and Sandra Carberry. Interpreting and generating indirect answers. *Computational Linguistics*, 25(3):389–435, 1999.

Herbert P Grice. Logic and conversation. In *Speech acts*, pages 41–58. Brill, 1975.

Jeroen Groenendijk and Floris Roelofsen. Compliance. In Alain Lecomte and Samuel Tronçon, editors, *Ludics, Dialogue and Interaction*, pages 161–173. Springer-Verlag, Berlin Heidelberg, 2011.

Jeroen Groenendijk and Martin Stokhof. Questions. In Johan van Benthem and Alice ter Meulen, editors, *Handbook of Logic and Linguistics*. North Holland, Amsterdam, 1997.

Jacob Hoepelmann. On questions. In Ferenc Kiefer, editor, *Questions and Answers*. Reidel, 1983.

Eduard Hovy, Ulf Hermjakob, and Deepak Ravichandran. A question/answer typology with surface text patterns. In *Proceedings of the second international conference on Human Language Technology Research*, pages 247–251. Morgan Kaufmann Publishers Inc., 2002.

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

Jacqueline C. Kowtko and Patti J. Price. Data collection and analysis in the air travel planning domain. In *Proceedings of the Workshop on Speech and Natural Language*, HLT '89, pages 119–125, Stroudsburg, PA, USA, 1989. Association for Computational Linguistics. ISBN 1-55860-112-0. doi: 10.3115/1075434.1075455. URL http://dx.doi.org/10.3115/1075434.1075455.

M. Krifka. For a structured meaning account of questions and answers. *Audiatur Vox Sapientia. A Festschrift for Arnim von Stechow*, 52:287–319, 2001.

Klaus Krippendorff. Agreement and information in the reliability of coding. *Communication Methods and Measures*, 5(2):93–112, 2011.

Theo AF Kuipers and Andrzej Wiśniewski. An erotetic approach to explanation by specification. *Erkenntnis*, 40(3):377–402, 1994.

Staffan Larsson. Comparing bdi approaches with the qud model. In J. Hulstijn and A. Nijholt, editors, *Proceedings of TwenDial 98, 13th Twente workshop on Language Technology*. Twente University, Twente, 1998.

Staffan Larsson. *Issue based Dialogue Management*. PhD thesis, Gothenburg University, 2002.

Alex Lascarides and Nicholas Asher. Agreement, disputes and commitments in dialogue. *Journal of Semantics*, 26(2):109–158, 2009.

Paweł Łupkowski and Olivia Ignaszak. Inferential erotetic logic in modelling of cooperative problem solving involving questions in the questgen game. *Organon F*, 24(2):214–244, 2017. URL `http://www.klemens.sav.sk/fiusav/doc/organon/2017/2/214-244.pdf`.

Paweł Łupkowski and Andrzej Wiśniewski. Turing Interrogative Games. *Minds and Machines*, 21 (3):435–448, Aug 2011. doi: 10.1007/s11023-011-9245-z. URL `http://dx.doi.org/10.1007/s11023-011-9245-z`.

Paweł Łupkowski, Mariusz Urbański, Andrzej Wiśniewski, Wojciech Błądek, Agata Juska, Anna Kostrzewa, Dominika Pankow, Katarzyna Paluszkiewicz, Oliwia Ignaszak, Joanna Urbańska, et al. Erotetic reasoning corpus. a data set for research on natural question processing. *Journal of Language Modelling*, 5(3):607–631, 2018.

Paweł Łupkowski and Jonathan Ginzburg. A corpus-based taxonomy of question responses. In *Proceedings of the 10th International Conference on Computational Semantics (IWCS 2013)*, pages 354–361, Potsdam, Germany, March 2013. Association for Computational Linguistics.

Paweł Łupkowski and Jonathan Ginzburg. Query responses. *Journal of Language Modelling*, 4(2): 245–293, 2016.

Brian MacWhinney. *The CHILDES project: Tools for analyzing talk*. Lawrence Erlbaum Associates, Mahwah, NJ, third edition, 2000.

Vladislav Maraev, Jonathan Ginzburg, Staffan Larsson, Ye Tian, and Jean-Philippe Bernardy. Towards KoS/TTR-based proof-theoretic dialogue management. In Lauren Prevot, Magali Ochs, and Benoit Fabre, editors, *Proceedings of SemDial 2018*, Aix-en-Provence, 2018.

Vladislav Maraev, Jean-Philippe Bernardy, and Jonathan Ginzburg. Dialogue management with linear logic: the role of metavariables in questions and clarifications. *Traitement Automatique des Langues (TAL)*, 61(3):43–67, 2020.

Per Martin-Löf. *Intuitionistic Type Theory*. Bibliopolis, Naples, 1984.

Mary L McHugh. Interrater reliability: the kappa statistic. *Biochemia medica*, 22(3):276–282, 2012.

Richard Montague. Pragmatics. In Richmond Thomason, editor, *Formal Philosophy*. Yale UP, New Haven, 1974.

Ira Noveck. *Experimental pragmatics: The making of a cognitive science*. Cambridge University Press, 2018.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peer Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Edouard Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12: 2825–2830, 2011.

Massimo Poesio and Hannes Rieser. (prolegomena to a theory of) completions, continuations, and coordination in dialogue. *Dialogue and Discourse*, 1:1–89, 2010.

Matthew Purver. *The Theory and Use of Clarification in Dialogue*. PhD thesis, King's College, London, 2004.

Matthew Purver. Clarie: Handling clarification requests in a dialogue system. *Research on Language & Computation*, 4(2):259–288, 2006.

Piotr Pęzik. Spokes search engine for Polish conversational data, 2014. URL `http://hdl.handle.net/11321/47`. CLARIN-PL digital repository.

Aarne Ranta. *Type Theoretical Grammar*. Oxford University Press, Oxford, 1994.

Craige Roberts. Information structure in discourse: Towards an integrated formal theory of pragmatics. *Working Papers in Linguistics-Ohio State University Department of Linguistics*, pages 91–136, 1996. Reprinted in Semantics and Pragmatics, 2012.

Carolyn P. Rosé, Barbara Di Eugenio, and Johanna D. Moore. A dialogue-based tutoring system for basic electricity and electronics. In Susanne P. Lajoie and Martial Vivet, editors, *Artificial intelligence in education*, pages 759–761. IOS, Amsterdam, 1999.

Emanuel Schegloff. *Sequence Organization in Interaction*. Cambridge University Press, Cambridge, 2007.

John R Searle. Indirect speech acts. In *Speech acts*, pages 59–82. Brill, 1975.

Asad Ali Shah, Sri Devi Ravana, Suraya Hamid, and Maizatul Akmar Ismail. Accuracy evaluation of methods and techniques in web-based question answering systems: a survey. *Knowledge and Information Systems*, 58(3):611–650, 2019.

Gabriel Skantze, Jens Edlund, and Rolf Carlson. Talking with higgins: Research challenges in a spoken dialogue system. In *International Tutorial and Research Workshop on Perception and Interactive Technologies for Speech-Based Systems*, pages 193–196. Springer, 2006.

Robert C. Stalnaker. Assertion. In P. Cole, editor, *Syntax and Semantics, Volume 9*, pages 315–332. AP, New York, 1978.

Anna-Brita Stenstrom. Questions and answers in English conversation. *Lund Studies in English, Malmo: Liber Forlag*, 1984.

Tanya Stivers. An overview of the question–response system in american english conversation. *Journal of Pragmatics*, 42(10):2772–2781, 2010.

Tanya Stivers. How we manage social relationships through answers to questions: The case of interjections. *Discourse Processes*, 56(3):191–209, 2019.

Tanya Stivers and Nick J Enfield. A coding scheme for question–response sequences in conversation. *Journal of Pragmatics*, 42(10):2620–2626, 2010.

Tanya Stivers and Jeffrey D Robinson. A preference for progressivity in interaction. *Language in society*, 35(3):367–392, 2006.

Tanya Stivers, Nicholas J Enfield, and Stephen C Levinson. Question-response sequences in conversation across ten languages: an introduction. *Journal of Pragmatics*, 42:2615–2619, 2010.

Ye Tian, Claire Beyssade, Yannick Mathieu, and Jonathan Ginzburg. Editing phrases. In Chris Howes and Staffan Larsson, editors, *Proceedings of GoDial, the 18th Workshop on the Semantics and Pragmatics of Dialogue*, Gothenburg, 2015. Gothenburg University.

A.M. Turing. Computing machinery and intelligence. *Mind*, 59(236):433–460, 1950.

Robert van Rooy. Asking to solve decision problems. *Linguistics and Philosophy*, 26(6):727–763, 2003.

Wei Wang. Grammatical conformity in question-answer sequences: The case of meiyou in mandarin conversation. *Discourse Studies*, pages 610–631, 2020.

Andrzej Wiśniewski. *Questions, Inferences, and Scenarios*. College Publications, London, England, 2013.

Andrzej Wiśniewski. Semantics of questions. In Chris Fox and Shalom Lappin, editors, *Handbook of Contemporary Semantic Theory, second edition*, Oxford, 2015. Blackwell.

Kyung-Eun Yoon. Questions and responses in korean conversation. *Journal of Pragmatics*, 42(10): 2782–2798, 2010.

Zulipiye Yusupujiang and Jonathan Ginzburg. Designing a GWAP for collecting naturally produced dialogues for low resourced languages. In *Workshop on Games and Natural Language Processing*, pages 44–48, Marseille, France, May 2020. European Language Resources Association. ISBN 979-10-95546-40-5. URL `https://www.aclweb.org/anthology/2020.gamnlp-1.7`.

Zulipiye Yusupujiang and Jonathan Ginzburg. Ugchdial: A uyghur chat-based dialogue corpus for response space classification. In *Proceedings of LREC 2022*, Marseille, France, 2022.

Zulipiye Yusupujiang, Alafate Abulmiti, and Jonathan Ginzburg. Classifying the response space of questions: A machine learning approach. In *Proceedings of SemDial 2022*, Dublin, Ireland, 2022.

Appendix A: Basic Notions of TTR

Type Theory with Records (Cooper, 2012; Cooper and Ginzburg, 2015; Cooper, 2023) is a cognitively construable formalism grounded in set theory, deriving much of its initial inspiration from Situation Semantics (Barwise and Perry, 1983; Ginzburg, 2011) and its formal notions from Constructive Type Theory (Martin-Löf, 1984; Ranta, 1994). A fundamental notion of Constructive Type Theory is the judgement $a : T$ that classifies an object a as being of type T. Failure to classify a by T is designated $a :\not T$.

Subtyping is defined as follows:

(59) $T_1 \sqsubseteq T_2$ iff $s : T_1$ implies $s : T_2$

Objects can be classified by relatively simple types such as those in (60);

(60) a. *Basic types* (*BType*; 0-place; *Ind*, *Loc*, *Time*, ...).

b. *Predicate types* (*PType*; n-place; lion(x), carry(x,y), ...), constructed out of a predicate and objects which are arguments of the predicate.

To classify more complex entities, for instance enable indefinite description, TTR introduces records and record types. A record is a set of fields assigning entities to labels of the form (61a), partially ordered by a notion of *dependence* between the fields on which their values depend. A concrete instance is exemplified in (61b). Records are used to model events and states, including utterances and dialogue gameboards.[36]

(61) a. $\begin{bmatrix} l_1 = val_1 \\ l_2 = val_2 \\ \dots \\ l_n = val_n \end{bmatrix}$

b. $\begin{bmatrix} \text{x} & = & 23 \\ \text{e-time} & = & \text{2AM, Sept 17, 1915} \\ \text{e-loc} & = & \text{kamen–kashirskiy} \\ \text{c}_{temp-at-in} & = & \text{o1} \end{bmatrix}$

A record type is simply a record where each field represents a judgement rather than an assignment, as in (62).

(62) $\begin{bmatrix} l_1 : T_1 \\ l_2 : T_2 \\ \dots \\ l_n : T_n \end{bmatrix}$

[36] Cooper and Ginzburg (2015) suggest that for events with even a modicum of internal structure, one can enrich the type theory using the "String theory" developed by Tim Fernando (e.g., (Fernando, 2007)).

The basic relationship between records and record types is that a record r is of type RT if each value in r assigned to a given label l_i satisfies the typing constraints imposed by RT on l_i. More precisely,

(63) The record
$$\begin{bmatrix} l_1 & = & a_1 \\ l_2 & = & a_2 \\ \dots & & \\ l_n & = & a_n \end{bmatrix} \text{ is of type } \begin{bmatrix} l_1 & : & T_1 \\ l_2 & : & T_2 \\ \dots & & \\ l_n & : & T_n \end{bmatrix}$$
iff $a_1 : T_1, a_2 : T_2, \dots, a_n : T_n$.

To exemplify this, (64a) (*the temperature of a given location at a given time*) is a possible type for (61b), assuming the conditions in (64b) hold. Record types are used to model utterance types (Saussurean/Formal Grammar *signs*) and to express rules of conversational interaction.

(64) a.
$$\begin{bmatrix} \text{x} & : & \text{Ind} \\ \text{e-time} & : & \text{Time} \\ \text{e-loc} & : & \text{Loc} \\ \text{c}_{temp-at-in} & : & \text{temp\_at\_in(e-time, e-location, x)} \end{bmatrix}$$

b. 23 : Ind; 2AM, Sept 17, 1915 : Time; kamen–kashirskiy : Loc; o1 : temp_at_in(2AM, Sept 17, 1915, kamen–kashirskiy, 23)

Sometimes one needs to partially specify a general type by tying down one or more of the fields to a specific value. For this we use a *manifest field* as in (65):

(65)
$$\begin{bmatrix} \text{x} & : & \textit{Ind} \\ \text{y=fido} & : & \textit{Ind} \\ \text{e} & : & \text{hug(x,y)} \end{bmatrix}$$

This is the type of situation where some individual hugs the individual 'fido'. Any record of this type would be one meeting the conditions in (66): .

(66)
$$\begin{bmatrix} \text{x} & = & a \\ \text{y} & = & b \\ \text{e} & = & s \\ \dots & & \end{bmatrix}$$
where a : *Ind*
b : *Ind* and b is 'fido'
s : hug(a,b)

TTR assumes in addition the following type construction operations:

(67) a. **Function types**: $(T_1)\text{T}_2$ is the type of functions from elements of type T_1 to type T_2.

b. **Set and list types**: *Set*(T) and *List*(T).

c. **Boolean types**: (i) Given a type T, there exists $\neg T$.
(ii) Given a set X of types T_i, there exist $\bigvee_X T_i$ and $\bigwedge_X T_i$.

$\bigvee_X T_i$ and $\bigwedge_X T_i$ have "classical" witnessing conditions:

(68) a. $r : \bigvee_X T_i$ iff for at least one $i \in X$ $r : T_i$

b. $r : \bigwedge_X T_i$ iff for all $i \in X$ $r : T_i$

In contrast, negation is a notion based on incompatibility that is a classical-intuitionist hybrid:

(69) a. $a : \neg T$ iff there is some T' such that $a : T'$ and T' *precludes* T

b. T' *precludes* T iff:

- $T = \neg T'$, or
- T and T' are non-negative and there is no a such that $a : T$ and $a : T'$

One can show that T and $\neg\neg T$ are equivalent, but the former is a positive, the latter a negative type. On the other hand, a need not be of type T and there need not be a type T' that precludes T; in other words: $a : T \vee \neg T$ is not a tautology. The basic reasoning for this goes back to (Barwise and Perry, 1983):

(70) a. If I observe Jo cutting onions, the situation I observe neither tells me that B. Johnson is smoking a cigar, nor that he is not smoking a cigar.

b. Hence, $\mathrm{s}_{visual} : Cutting(j, o)$, $\mathrm{s}_{visual} \not: CigarSmoke(b.johnson)$, hence: it is not the case that $\mathrm{s}_{visual} : CigarSmoke(b.johnson)$, but neither is it the case that $\mathrm{s}_{visual} : \neg CigarSmoke(b.johnson)$

The final notion we mention are *propositions*.[37] Propositions are construed as typing relations between records (situations) and record types (situation types), or Austinian propositions (Austin, 1961; Barwise and Etchemendy, 1987); more formally:

(71) a. Propositions are records of type
$$\text{Prop} = \begin{bmatrix} \text{sit} & : & \textit{Rec} \\ \text{sit-type} & : & \textit{RecType} \end{bmatrix}.$$

b. $\text{p} = \begin{bmatrix} \text{sit} & = & \text{s} \\ \text{sit-type} & = & \text{T} \end{bmatrix}$ is true iff *p.sit* : *p.sit-type* i.e., $s : T$ —the situation s is of the type T.

Two subtypes of Austinian propositions are given in (72b,c):

[37] For a TTR approach using solely types and for detailed discussion of the two approaches, see Chapter 6 of (Cooper, 2023).

(72) a. $Sign = \begin{bmatrix} \text{phon} & : & List(Phonform) \\ \text{cat} & : & \left[\text{head} \quad : \quad PoS\right] \\ \text{dgb-params} & : & RecType \\ \text{q-params} & : & RecType \\ \text{cont} & : & SemObj \end{bmatrix}$

b. For classifying utterances, as described in the text:

$Loc(utionary)Prop(osition) = \begin{bmatrix} \text{sit} & : & Rec \\ \text{sit-type} & : & Sign \end{bmatrix}$

c. For assigning contents to dialogue moves:

$Illoc(utionary)Prop(osition) = \begin{bmatrix} \text{sit} & : & Rec \\ \text{x} & : & Ind \\ \text{y} & : & Ind \\ \text{a} & : & Prop \vee Question \\ \text{R} & : & IllocRel \\ \text{sit-type} & = & \left[\text{c1} \quad : \quad \text{R(x,y,a)}\right] : RecType \end{bmatrix}$

Appendix B: Annotation guidelines

Below we present the full annotation guidelines (in English and in Polish) used in the described corpus study. The alert reader will notice that the number of question responses categories in the guidelines is larger than the number of categories discussed in the paper (see Figure 1). The reason for this is that we decreased the number of categories initially posited by merging selected ones. The motivation for this move comes from the analysis of the annotation reliability and disagreement cases. We decided to merge (i) IA into IND, (ii) FORM and COR into CR, and (iii) IDK into DPR. As a result, we have more general categories and we avoid a situation where we have categories with only few cases present in our data.

We provide additional data for this paper (covering annotated Q-R pairs and disagreement cases) which are hosted on the OSF web-page (`https://osf.io/mq6r7/`).

Annotation guidelines
Instrukcja dla anotatorów

Is the utterance an ANSWER (provides information required by a question) or a NON-ANSWER.
Czy reakcja na pytanie jest ODPOWIEDZIĄ (dostarcza informacji wymaganych przez pytanie) czy NIE-ODPOWIEDZIĄ?

If ANSWER, then
Jeżeli ODPOWIEDŹ, to
DA = direct ANSWER (provides the required information straightforwardly).
A: Who is going to check that? / B: I can check that.
A: and how long did they sleep? long? / B: well, you know, Stas slept for at least two hours
DA = odpowiedź bezpośrednia (dostarcza wymaganych informacji wprost)
A: Kto to sprawdzi? / B: Ja mogę to sprawdzić.
A: a długo spali? długo spali? / B: wiesz co no Staś to ze dwie godziny spał

IA = indirect ANSWER (you need to infer an answer from the utterance, it is not straightforward).
A: Do you want more tapes for them to take away? / B: I've got ten. I've haven't used any of them.
IA = odpowiedź pośrednia (wymagane jest wywnioskowanie odpowiedzi z wypowiedzi, nie jest ona podana wprost)
A: Chcesz więcej taśm, żeby zabrać je ze sobą? / B: Mam dziesięć. Nie użyłem żadnej z nich.

Else: NON-ANSWER, then
Is it a QUESTION? If QUESTION, then
Jeżeli NIE-ODPOWIEDŹ, to:
Czy jest to PYTANIE? Jeżeli PYTANIE, to:
CR = Is q2 a query about something not completely understood in q1?
A: Why are you in? / B: What?
CR = q2 jest zapytaniem o coś nie do końca zrozumianego w q1, prośba o wyjaśnienie
A: Dlaczego jesteś w środku? / B: Co?
A: na pewno a jest już? / B: proszę?

DP = Is it the case that the answer to q1 depends on the answer to q2?
A: Do you want me to <pause> push it round? / B: Is it really disturbing you?
DP = przypadek, w którym odpowiedź na q1 zależy od odpowiedzi na q2
A: Czy chcesz żebym <pauza> popchnął to dookoła? / B: Czy naprawdę Ci to przeszkadza?

IGNORE = Does q2 relate to the situation described by q1?
A: Just one car is it there? / B: Why is there no parking there?
IGNORE = q2 odnosi się do sytuacji opisanej w q1, natomiast nie pośrednio do q1
A: Tam jest tylko jeden samochód? / B: Czemu tam nie ma parkingu?
A: a i był ten merlot co w Łodzi żeśmy pili to bardzo dobry był nie? / B: czternaście złotych chyba on kosztował?

IND = Is it the case that q2 is rhetorical and in this sense does not need to be answered and provides (indirectly) an answer to q1?
A: What is it? / A: What's he done? / B: Ehm, you know what I've said before, eh, eh you'll get <unclear>
IND = przypadek, w którym q2 jest retoryczne, nie musi uzyskać odpowiedzi i dostarcza (pośrednio) odpowiedzi na q1
A: Co jest? / A: Co on zrobił? / B: Hmm, wiesz, to co powiedziałem wcześniej, hm, hm, dostaniesz <niejasne>

FORM = Is it the case that the way the answer to q1 will be given depends on the answer to q2?
A: Okay then, Hannah, what, what happened in your group? / B: Right, do you want me to go through every point?
FORM = przypadek, w którym sposób w jaki odpowiedź na q1 będzie wyglądała zależy od odpowiedzi na q2
A: Dobrze więc, Hannah, co, co się stało w Twojej grupie? / B: Dobrze, czy chcesz żebym przeszła przez każdy punkt?

MOTIV = Does q2 address the motivation underlying asking q1?
A: What's the matter? / B: Why?
MOTIV = q2 pyta o motywację leżącą u podstaw q1
A: Co się stało? / B: Dlaczego pytasz?

Is it a DECLARATIVE? If DECLARATIVE
Czy jest to DEKLARATYW (zdanie twierdzące)? Jeżeli DEKLARATYW, to:

IDK = I do not know, the speaker states that s(he) does not know the answer
A: When's the first consignment of Scottish tapes? / B: Erm <pause> don't know.
IDK = mówca daje do zrozumienia, iż nie zna odpowiedzi
A: Kiedy to było? / B: Erm <pauza> nie wiem.
A: to jest Agnieszki ten koleś? / B: nie wiem ale ciszej

DPR = difficult to provide an answer, the speaker states that it is hard to provide response, points at a different information source, etc.
A: Why? / B: i'm not exactly sure.
DPR = trudność w podaniu odpowiedzi, mówca oświadcza, że podanie odpowiedzi jest trudne, wskazuje na inne źródło informacji, itd.
A: Czemu? / B: Nie jestem pewien.

COR = correction, the speaker point that a question has a wrong presupposition
A: UB forty? / B: WD forty.
COR = deklaratywny odpowiednik CR, zakłada coś związanego z intencjami oryginalnego mówcy zamiast pytać, odpowiedź wskazująca na błędne założenie obecne w pytaniu
A: UB forty? / B: WD forty.

ACK = Acknowledgement, a speaker letting know that s(he) heard the question, e.g. mhm, aha etc.
A: that's about it innit? / B: Mm mm.
ACK = potwierdzenie, mówca daje znać iż usłyszał/a pytanie poprzez na przykład mhm, aha itd.
A: Czy to już wszystko? / B: Mhm.
A: wiesz jak ma na imię? / B: poczekaj

CHT = an utterance that signals that the speaker does not want to answer, s(he) changes the topic, provides an evasive answer.
A: What's dolly's name? / B: It's raining.
CHT = wypowiedź sygnalizuje iż mówca nie chce odpowiedzieć, zmienia temat, udziela odpowiedzi wymijającej
A: Jak Dolly się nazywa? / B: Deszcz pada.
A: czarne podoba ci się ? / B: brud widać

IGNORE = the utterance does not relate to the question, but to the situation
A: So does that mean that the ammeter is not part of the series, just hooked up after to the tabs? / B: Let's take a step back.
A: What have you been doing Melvin? <laugh> / B: I ain't talking cos you've got that bloody thing on
IGNORE = reakcja odnosi się do sytuacji opisanej w q1, natomiast nie pośrednio do q1
A: Melvin co Ty robiłeś? <śmiech> / B: Nic nie powiem bo masz to coś na sobie.
A: ale w jakim pokoju? / B: no wiesz że on tam wiesz wyje trochę się uspokaja potem znowu wyje no

In all other cases, put the OTHER tag.
W innych przypadkach, proszę użyć tagu OTHER a w kolumnie obok opisać jaką funkcję spełnia ta reakcja na pytanie w tym konkretnym przypadku.