Characterizing the Response Space of Questions: a Corpus Study for English and Polish

Jonathan Ginzburg Zulipiye Yusupujiang Chuyuan Li Kexin Ren
Université de Paris, CNRS, Laboratoire de Linguistique Formelle
yonatan.ginzburg@univ-paris-diderot.fr

Paweł Łupkowski
Adam Mickiewicz University, Poznań
Pawel.Lupkowski@amu.edu.pl

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). This work develops a wide coverage taxonomy for question/question sequences observable in corpora including the BNC, CHILDES, and BEE, as well as formal modelling of all the postulated classes. Our aim is to extend this work to cover *all* responses to questions. We present the extended typology of responses to questions based on a corpus studies of BNC, BEE and Maptask with include 506, 262, and 467 question/response pairs respectively. We compare the data for English with data from Polish using the Spokes corpus (205 question/response pairs). We discuss annotation reliability and disagreement analysis. We sketch how each class can be formalized using a dialogical semantics appropriate for dialogue management.

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 queries.

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 (Turing, 1950).

(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. 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 question responses.[1] We take their work as a starting point and make the following hypothesis:

(1) Main hypothesis: responses drawn from or concerning the LG classes plus direct and indirect 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 queries (A: Who should we invite? B: Who is in town?);
2. Clarification Requests.
3. Evasion responses:
   (a) Ignore (address the situation, but not the question);
   (b) Change the topic ('Answer *my* question');
   (c) Motive ('Why do you ask?');
   (d) IDK ('I don't know');

[1] The study sample consisted of 1,466 query/query response pairs. As an outcome the following query responses (q-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 an underlying motivation behind asking the initial question; (4) NO ANSW: questions aimed at avoiding answering the initial question; (5) FORM: questions considering the way of answering the initial question; (6) QA: questions with a presupposed answer, (7) IGNORE: responses ignoring the initial question—for more details see (Łupkowski and Ginzburg, 2016, p. 355).

Proceedings of the SIGDial 2019 Conference, pages 320–330
Stockholm, Sweden, 11-13 September 2019. ©2019 Association for Computational Linguistics

(e) 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, as exemplified in (2), 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 provide a brief discussion of the existing literature in section 2. Following this, we provide a description of the proposed taxonomy, in section 3. We then set out to test our main hypothesis in an initial study, using three corpora in English (BNC, BEE, MapTask) and one corpus in Polish (Spokes (Pezik, 2015)). By and large, the hypothesis achieves wide coverage, as we discuss in section 5. We sketch an account of how the different classes can be characterized, taking a fairly general perspective and building on the initial characterization of (Łupkowski and Ginzburg, 2016) while drawing some metatheoretical conclusions. Finally, section 8 offers a variety of extensions we plan to undertake.

2 Related work

Berninger and Garvey (1981) introduce their rich taxonomy of possible replies for children conversation in a nursery school. The taxonomy covers six categories, categories that are co–extensive with the ones mentioned in the introduction to this paper, though no semantic explication or interannotator study is offered: (i) Indirect answers. (ii) Confessions of ignorance. (iii) Clarification questions. (iv) Evasive replies. (v) Miscellaneous.

An extensive 10-language comparative project on question/response sequences in ordinary conversation was carried out from 2007 as the part of the Multimodal Interaction Project at the Max Planck Institute for Psycholinguistics (Stivers et al., 2010). The coding scheme for the response types covered categories of **Non-response**, **Non-answer response**, **Answer**, and **Can't determine** (Stivers and Enfield, 2010, p. 2624).

The results were 76% answer responses, 19% non-answers, and 5% non-responses. (Stivers, 2010, p. 2778) 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). It is worth stressing that the question sample was limited to questions that functionally sought information, confirmation or agreement see (Yoon, 2010, p. 2783).

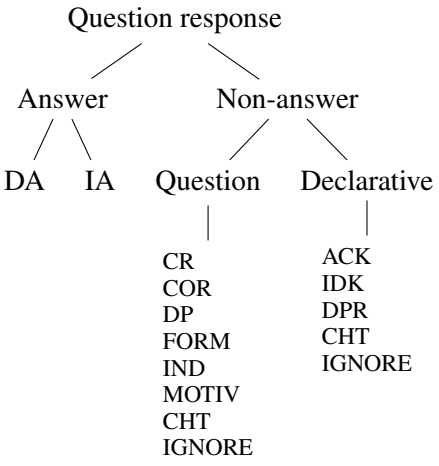

Figure 1: Response space of questions

The work discussed in this section indicate the need for a wider corpus study of the whole spectrum of answers to questions.[2] The studies discussed are limited in terms of analyzed examples. They also imposed certain limitations in terms of numbers of response categories to be identified—they were mainly aimed at understanding the answer/non-answer difference. An extensive corpus study is needed for a fine grained characterization of the response space of questions. Moreover, we aim at providing an explicit dialogical semantics for each category of our corpus-based typology.

3 A taxonomy of responses to queries

We start with the most general division of question responses to answers and non-answers as discussed in the previous section. In the answer class we distinguish direct and indirect answers—see figure 1.

[2]For a detailed review of the literature on query responses, see (Łupkowski and Ginzburg, 2016),pp. 245-49, which discusses work from the question generation literature, in particular (Graesser et al., 1992).

Direct answers (DA) are (i) either sentential and denote propositions that are answers or (ii) are non-sentential and convey an answer as their content.[3] This is clearly visible in the following example—B is providing information required by A:

(3) A: Who is going to check that?
B: *Well I can check it.*

Indirect answers (IA) involve an inference of an answer from the utterance, as in (4):[4]

(4) A: What is it?
A: What's he done?
B: *Ehm, you know what I've said before.*

Here A has to infer the answer to his/her questions from B's suggestion that this issue has been addressed before.

For the non-answer group the taxonomy (mostly) reuses the classes proposed in (Łupkowski and Ginzburg, 2013, 2016) with some minor renaming.

Clarification questions (CR) address something that was not completely understood in initial question (q1)[5], like:

(5) A: Why are you in?
B: *What?*

Corrections (**COR**) are declarative counterparts of CRs in that they assert rather than query about the original speaker's intended meaning. This is exemplified in B's answer in (6):

(6) A: what is it?
A: Something forty <unclear>.
A: UB forty?
B: *WD forty.*
A: WD.

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:

(7) 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.*]

See more in section 7.1.

Question responses may also address that the way the answer to q1 will be given depends on the answer to q2 (**FORM**). This type of question response differs from DP as the response concerns only the form in which the answer to q1 will be given (how it will be formulated). This may be noticed in (8), where the way B answers A's question will be dictated by A's answer to q2—whether or not A wants to know details point by point.

(8) A: Okay then, Hannah, what, what happened in your group?
B: *Right, do you want me to go through every point?*

One also encounters q2, which is rhetorical and in this sense does not need to be answered and **indirectly provides an answer** to q1 (IND).

(9) A: Are you Gemini?
B: *Well if I'm two days away from your, what do you think?*

As for evasive question-responses we have one 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 the following example:

(10) A: What's the matter?
B: *Why?*

Another type of evasive question-response is **change–the–topic (CHT)**. These are cases wherein q2 enables the speaker to avoid answering q1 while attempting to force the other speaker to answer q2 first. Instead of answering q1, the agent provides q2 and attempts to "turn the table" on the original querier. The original querier is pressured to answer q2 and put q1 aside.

[3] For the direct answers category we allow for additional sub-categories, which we return to discuss briefly in section 7. These include: (1) no/yes answer to polar questions; (2) simple answer to wh-questions; (3) partial polar answer; (4) partial wh-question answer.

[4] As with the direct answers category, we have also used the following sub-categories of indirect answers, but do not elaborate on this here for reasons of space: (i) indirect answer addressing wh-question; (2) q-widening IAs (over-informative answer to a polar question, addressing a more general wh-question).

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

(11) A: Why is it recording me?

B: *Well why not?*

An **IGNORE** type of query-response appears when q2 relates to the situation described by q1 but not directly to the initial question:

(12) 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?*

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). See also the following example:

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

B: *Why is there no parking there?*

If a question response is not an answer and it is a declarative we consider the following cases. For a start declarative responses can serve the same purpose as ignoring query-response:

(14) a. 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.*

b. A: What have you been doing Melvin? <laugh>

B: *I ain't talking cos you've got that bloody thing on.*

Acknowledgement (ACK)—a speaker acknowledges that s(he) has heard the question, e.g. *mhm*, *aha* etc.

(15) A: that's about it innit?

B: *Mm mm.*

The speaker states that s(he) **does not know the answer** (IDK).

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

B: *Erm <pause> don't know.*

The speaker states that it is **hard to provide an answer** (DPR), points at a different information source, etc.

(17) A: Why?

B: *I'm not exactly sure.*

An utterance signalizes that speaker does not want to answer, s(he) **changes the topic**, gives an evasive answer (CHT).[6]

(18) A: What's dolly's name?

B: *It's raining.*

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

The data for English comes from the BNC, BEE, and the MapTask corpora (Burnard, 2000; Rosé et al., 1999; Anderson et al., 1991). 506 Q-R turns were taken from the BNC, 256 Q-R turns from BEE, and 467 Q-R turns from the MapTask. In each case starting points where questions occur were chosen by randomly selecting turn numbers, and coding the subsequent questions in that extract. Questions were turn units ending with a '?'; however, tag questions and turns with missing text (the BNC's 'unclear') were eliminated from considerations. The BNC data covers mainly topically unrestricted conversations. As for BEE and MapTask dialogues are more task oriented—BEE contains contains tutorial dialogues from electronics courses and MapTask consists of dialogues recorded for a direction–providing task.

4.2 Polish: the Spokes Corpus

The data used for this study was drawn from the Spokes corpus (Pezik, 2015). The corpus currently contains 247,580 utterances (2,319,291 words) in

[6] 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 institutionalisation of anti-Semitism among other progressives? Sadly, my answer to all of these is an unequivocal yes. (D Lipstadt, *Antisemitism: Here and Now*)

transcriptions of spontaneous conversations. For the study four files were selected from the corpus (10,244 words, 1,424 turns)[7]. Within each file the question-response pairs (Q-R) were selected manually. In total we obtained 205 Q-R pairs for the study.

5 Results

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. Also the OTHER category was included. The tagset used to annotate gathered data is presented in Table 1. The detailed results of the annotation are presented in figure 2. We discuss the annotation reliablity in section 6.

5.1 English

In all three cases, the OTHER class is less than 3%, hence coverage is above 97%. The most frequent classes of responses in all three corpora are direct answers (DA); in the BNC the next biggest are clarification requests, for BEE these are indirect answers, whereas for the MapTask the second biggest are IGNORE.

5.2 Polish

The two most frequent classes of responses for Spokes are answers: direct ones (DA=51.71%) and—much smaller—indirect ones (IA=13.66%).The next two most frequent classes are IDK (stating that a person does not know the answer to the question, IDK=10.24%) and utterances ignoring the question asked (questions and declaratives, IGNORE=9.76%).

5.3 Discussion

As might be expected from the results presented in (Łupkowski and Ginzburg, 2016), the most frequent *question*-response for English and Polish data is the clarification request. What is more surprising is that by adding declaratives into the picture a relatively high number of ignoring responses is observed for both English and Polish. Łupkowski and Ginzburg (2016) analyzed only question-responses and this type was observed rarely (0.57% for n=1,051 for BNC). Other evasive responses (relatively) frequent in both lan-

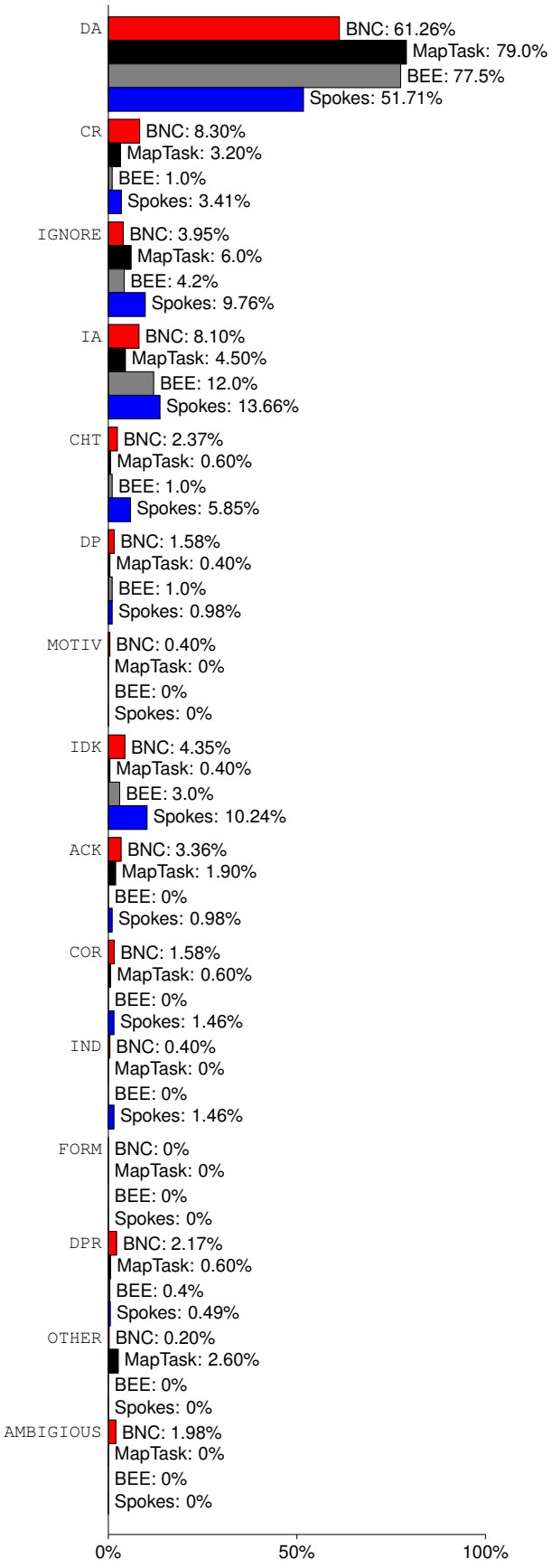

Figure 2: Frequency of responses to questions for the BNC (n=506), BEE (n=256), MapTask (n=467) and Spokes (n=205) studies

[7]Files 016O, 019w, 01AO, 01dL cover casual conversation concerning youth, wine and travelling plans.

guages are CHT and IDK. For the latter, we observe that it was more frequent in Polish than in the English data. This may be a consequence of the lower number of examples analyzed for Polish—Spokes is smaller and less varied than the BNC.

As regards cross-corpus differences, BNC and Spokes data cover mainly topically unrestricted conversations, while BEE and MapTask contain task-oriented dialogues. Correspondingly, MapTask has the highest number of direct answers (79.0%), and BEE almost the same (77.5%). However, for BNC and Spokes these numbers are lower (respectively 61.26% and 51.71%). For both clarification requests and evasive response types frequencies are lower for task-oriented corpora than for BNC and Spokes (this is in line with results for BNC and BEE reported in (Łupkowski and Ginzburg, 2016, p. 256–257)).

6 Annotation reliability

6.1 Inter-annotator studies

Table 1: Tagset used for annotation of the data

Category	TAG
1. Direct answer	DA
2. Indirect answer	IA
3. Clarification response	CR
4. Dependent question	DP
5. The utterance does not relate to the question, but to the situation	IGNORE
6. Question being an indirect answer	IND
7. Question addressing the form of answer to be given	FORM
8. Question about the motivation for the initial question	MOTIV
9. I do not know	IDK
10. Difficult to provide an answer	DPR
11. Correction	COR
12. Acknowledgement	ACK
13. Utterance signalizes that speaker does not want to answer, s(he) changes the topic, gives evasive answer	CHT
14. Utterance that does not fit in any of the above	OTHER

For English: For the inter-annotator study a sample of nearly 800 Q-Rs from the BNC were annotated by two advanced graduate students in computational linguistics, L2 speakers of English, who underwent several training sessions with one of the authors, a native speaker of English with significant experience in dialogue annotation. The first annotator coded 622 Q-Rs and the second annotator annotated 730 Q-Rs. Then we chose the initial 515 Q-Rs, which were commonly annotated by both annotators, deleting 9 Q-Rs which were incomplete or unclear utterances to yield the 506 commonly annotated QR pairs from the BNC. For these we calculated the κ (Carletta, 1996) and α (Krippendorff, 2011) measures. We used the data mining and data analysis tool (Pedregosa et al., 2011) in Python with its *sklearn.metrics* package for calculating Cohen's kappa, and also used the Python implementation *Krippendorff*[8] for the calculation of Krippendorff's alpha. In this case, Cohen's Kappa for two annotators is 0.65 (substantial), and Krippendorff's alpha is 0.66. All disagreements were then discussed in detail by one of the annotators and the afore-mentioned author and resolved (though some ambiguous cases remain, as discussed below.).

For Polish: The entire 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). Fleiss' Kappa for all three annotators was 0.53 (i.e. moderate). For the first and the second annotator—Cohen's Kappa 0.66 (substantial). For the first and the third annotator—Cohen's Kappa 0.49 (moderate).[9] Krippendorff's alpha for all three annotators is 0.742. For the first and second annotator the score is 0.617, while for the first and the third annotator it is 0.379. All measures were calculated using the *irr* package (Gamer et al., 2012) from R (R Core Team, 2013), version 3.3.1.

Disagreement analysis For reasons of space, we restrict attention to English here. Among the valid commonly annotated 506 BNC Q-Rs, there are 94 cases where the annotation disagreements between two annotators occurred The main disagreements concerned DA versus IA (34), IGNORE versus CHT/ACK/DP/DA (16), and ACK versus OTHER (5), as exemplified in (19). 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).

[8] https://pypi.org/project/krippendorff/

[9] Whereas the first and second annotators have much experience in dialogue annotation, the third annotator is a logician with less annotation experience.

(19) a. ANON5: Why do they pretend not to know?

ANON5: <pause> I mean they should be fully aware of of of our <unclear>

ANON2: **Val, well this is a new guy.** [*DA v. IA, resolved to IA.*]

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 IA since indirectly indicates that there is no reason.*]

c. JOHN: Can you spell box?

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

After carefully discussing all disagreements, we concluded that there are (at least) 10 cases which are truly ambiguous and should not be resolved; this is in line with a recent trend in dialogue annotation (e.g., Passonneau and Carpenter, 2014); though we have not implemented the more complex approach this inevitably requires in the current work. We exemplify two such cases. (20a,b) involve an ambiguity between CR and IND, and DA and IA, respectively; both are hard to resolve conclusively.

(20) a. FRANCIS: What is five?

FRANCIS: Tell me <unclear>.

UNKNOWN: **<pause> is there five people?**

b. HUG: What's he working on Rog?

ROG: **Oh he's off work <unclear> and you see he has all the time off for councils and you know it isn't as if he's there fulltime.**

7 Formal Analysis

In this section, we discuss briefly the requirements on a computational semantic theory to be able to characterize the response space of a query in terms of the notions discussed in previous sections. Łupkowski and Ginzburg (2016) assume such a characterization should be formulated in dialogical terms, for instance as dynamics of agent information states, since this makes the analysis usable for dialogue analysis. Indeed, to the extent that the empirical work here verifies our main hypothesis (1), the formal rules provided in (Łupkowski and Ginzburg, 2016) yield a complete characterization of the response space for questions in implementable form (for a sketch see (Maraev et al., 2018)). However, using a proof theoretic approach along the lines of erotetic logics like IEL (Wiśniewski, 2013) is conceivable, assuming it can be extended in certain respects, as we will explain.

7.1 Question–specificity

Any speaker of a given language can recognize, independently of domain knowledge and of the goals underlying an interaction, that certain propositions are *about* or *directly concern* a given question. This is the answerhood relation needed for characterizing direct answerhood.

The most basic notion of answerhood—*simple answerhood* (Ginzburg and Sag, 2000)—is the range of the propositional abstract, plus their negations.

(21) a. $\text{SimpleAns}(\lambda\{\ \}p) = \{p, \neg p\}$;

b. $\text{SimpleAns}(\lambda x.P(x)) = \{P(a), P(b), \ldots, \neg P(a), \neg P(b) \ldots\}$

In fact, *simple answerhood*, though it has good coverage, is not sufficient. *Aboutness* must be sufficiently inclusive to accommodate conditional, weakly modalized, and quantificational answers, all of which are pervasive in actual linguistic use (Ginzburg and Sag, 2000).

How to formally and empirically characterize aboutness is an interesting topic researched within work on the semantics of interrogatives (see e.g. Ginzburg and Sag, 2000; Groenendijk, 2006), though a comprehensive, empirically-based, experimentally tested account for a variety of wh–words is still elusive.

An additional important notion a theory of questions needs to provide for is a notion of *exhaustiveness*, though this is in general pragmatically parametrized (Asher and Lascarides, 2003). Whether a response is (pragmatically) exhaustive (or *goal fulfilling*) can determine whether the response will be accepted or require a follow up query. Hence, the need for a finer–grained subdivision of the answer categories, as we hinted in footnotes 3 and 4.

Given a notion of aboutness and some notion of (partial) exhaustiveness, one can then define question dependence (needed for the class DP), for instance, as in (22), though various alternative definitions have been proposed (Groenendijk and Stokhof, 1997; Wiśniewski, 2013; Onea, 2016). For all these definitions their coverage awaits testing on empirical data:

(22) $q1$ depends on $q2$ iff any proposition p such that p resolves $q2$, also satisfies p entails r such that r is about $q1$. (Ginzburg, 2012, (61b), p. 57)

With notions of aboutness and dependency in hand, one can define update rules licensing such responses. For instance, a rule of the following form:

(23) QSPEC: If q is the question under discussion, respond with an utterance r which is q–specific: About(r,q) or Depends(q,r)

7.2 Repair utterances

Clarification requests and (metacommunicative) corrections is a domain where logics that use simply contents of utterances are not adequate (Ginzburg and Cooper, 2004). Their generation requires access to the entire sign associated with a given interrogative utterance. (Purver, 2004; Ginzburg, 2012) show how to account for the main classes of CRs using rules that enable clarification questions relevant to a given utterance under clarification to be accommodated into the content. Each such rule specifies an accommodated MAX-QUD built up from a sub-utterance u1 of the target utterance, the maximal element of the Pending attribute of the context (*MAX-Pending*). Common to all these rules is a license to follow up *MAX-Pending* with an utterance which is *co-propositional* with MAX-QUD.[10] Abstracting away from formal details, such rules can be specified as in (24), with the three disjuncts indicating the possible clarification questions that can be accommodated:

(24) **Clarification Context Update Schema**
Input: u: utterance by A, u1, constituent of u
Output:
MAXQUD:
(i)reference resolution: *what did A mean by u1* ,
(ii)form resolution: *what word did A utter at u1*,
(iii)confirmation of constituent content: *what is u1's content x, given that u's content is C(x)*

7.3 Evasion Utterances

A natural way to analyze utterances relating to MOTIV is along the lines of a rule akin to QSPEC above: If A has posed q, B may follow up with an utterance specific to the issue *?Wish-Answer(B,q)*

(Łupkowski and Ginzburg, 2016) postulate fairly strong constraints on CHT and IGNORE to ensure that they are not unrestricted and do not allow any issue in. IGNORE is assumed to require the issue to be situationally shared with the posed question $q1$. This requires a means of evaluating shared–situatedness between questions. For CHT they assume that the topic changing question q2 introduced by or addressed by the response must be unifiable with q1 via a third question q3 (e.g., $q1$ = what do you (B) like? $q2$ = what do you (A) like? $q3$ = Who likes what?.). This requires a question inference mechanism for testing this unifiability.

8 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 queries. Achieving such a characterization is a fundamental challenge for semantics

[10] Two utterances are co-propositional if, modulo their domain, the questions they introduce into QUD involve similar answers—a query q introduces q into QUD, whereas an assertion p introduces $p?$ into QUD. For instance 'Whether Bo left', 'Who left', and 'Which student left' (assuming Bo is a student) are all co-propositional. Hence the available follow ups licensed in this way are clarification requests that differ from MAX-QUD at most in terms of its domain, or acknowledgements and corrections—propositions that instantiate MAX-QUD.

with a very wide variety of applications. It also establishes basic theoretical benchmarks for theories of dialogue/discourse 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:

- 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:
  1. The learnability of non sentential answers (Fernández et al., 2007; Dragone and Lison, 2015) gives hope for learnability of some other classes.
  2. On the other hand, we anticipate significant difficulty with learning heavily inference-based classes like indirect answers, and IGNORE/CHT.
- Spoken dialogue system implementation: we plan to test the usability of these categories in dialogue systems with sophisticated dialogue management (Larsson and Berman, 2016) and NLU (see Maraev et al., 2018).
- 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 et al., 2018).

Acknowledgments

We acknowledge the support of the French Investissements d'Avenir-Labex EFL program (ANR-10-LABX-0083) and a senior fellowship from the Institut Universitaire de France to the first author, which funded the internships of Yusupujiang, Li, and Ren at LLF. We thank three anonymous reviewers for SigDial for their detailed and perceptive comments. A much earlier version of this work was presented at the workshop *Why indeed? Questions at the interface of theoretical and computational linguistics* in Stuttgart in March 2018. Thanks to the organizers, Annette Hautli-Janisz, Aikaterini-Lida Kalouli und Tatjana Scheffler, and the audience for its feedback.

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. 1991. The HCRC Map Task Corpus. *Language and Speech*, 34(4):351–366.

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

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

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

L. Burnard. 2000. *Reference Guide for the British National Corpus (World Edition)*. Oxford Universtity Computing Services.

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

Paolo Dragone and Pierre Lison. 2015. An active learning approach to the classification of non-sentential utterances. In *Proceedings of the Second Italian Conference on Computational Linguistics*, pages 115–119.

Raquel Fernández, Jonathan Ginzburg, and Shalom Lappin. 2007. Classifying ellipsis in dialogue: A machine learning approach. *Computational Linguistics*, 33(3):397–427.

Matthias Gamer, Jim Lemon, and Ian Fellows Puspendra Singh. 2012. *irr: Various coefficients of interrater reliability and agreement*. Acess 20.03.2017, R package version 0.84.

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

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

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

A. C. Graesser, N. K. Person, and J. D. Huber. 1992. Mechanisms that generate questions. In T. E. Lauer, E. Peacock, and A. C. Graesser, editors, *Questions and information systems*, pages 167–187. Lawrence Erlbaum Associates, Hillsdale.

Jeroen Groenendijk. 2006. The logic of interrogation. In Maria Aloni, Alistair Butler, and Paul Dekker, editors, *Questions in Dynamic Semantics*, volume 17 of *Current Research in the Semantics/Pragmatics Interface*, pages 43–62. Elsevier, Amsterdam. An earlier version appeared in 1999 in the Proceedings of SALT 9 under the title 'The Logic of Interrogation. Classical version'.

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

Jacqueline C. Kowtko and Patti J. Price. 1989. 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. Association for Computational Linguistics.

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

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

Staffan Larsson and Alexander Berman. 2016. Domain-specific and general syntax and semantics in the talkamatic dialogue manager. *Empirical Issues in Syntax and Semantics*, 11:91–110.

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

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

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

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

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

Edgar Onea. 2016. *Potential questions at the semantics-pragmatics interface*. Brill, Leiden, Boston.

Rebecca J Passonneau and Bob Carpenter. 2014. The benefits of a model of annotation. *Transactions of the Association for Computational Linguistics*, 2:311–326.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Piotr Pezik. 2015. Spokes – a search and exploration service for conversational corpus data. In *Selected Papers from the CLARIN 2014 Conference, October 24-25, 2014, Soesterberg, The Netherlands*, 116, pages 99–109. Linköping University Electronic Press, Linköpings universitet.

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

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

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

R Core Team. 2013. *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. Acess 20.03.2017.

Carolyn P. Rosé, Barbara Di Eugenio, and Johanna D. Moore. 1999. 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.

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

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

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

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

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

Andrzej Wiśniewski. 2015. Questions. In *Handbook of Contemporary Semantic Theory, second edition*, Oxford. Blackwell.

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