# Context Modeling for IQA: The Role of Tasks and Entities

**Raffaella Bernardi** and **Manuel Kirschner**
KRDB, Faculty of Computer Science
Free University of Bozen-Bolzano, Italy
{bernardi, kirschner}@inf.unibz.it

## Abstract

In a realistic Interactive Question Answering (IQA) setting, users frequently ask follow-up questions. By modeling how the questions' focus evolves in IQA dialogues, we want to describe what makes a particular follow-up question salient. We introduce a new focus model, and describe an implementation of an IQA system that we use for exploring our theory. To learn properties of salient focus transitions from data, we use logistic regression models that we validate on the basis of predicted answer correctness.

## 1 Questions within a Context

Question Answering (QA) systems have reached a high level of performance within the scenario originally described in the TREC competitions, and are ready to tackle new challenges as shown by the new tracks proposed in recent instantiations (Voorhees, 2004). To answer these challenges, attention is moving towards adding semantic information at different levels. Our work is about context modeling for Interactive Question Answering (IQA) systems. Our research hypothesis is that a) knowledge about the dialogue history, and b) lexical knowledge about semantic arguments improve an IQA system's ability to answer follow-up questions. In this paper we use logistic regression modeling to verify our claims and evaluate how the performance of our Q→A mapping algorithm varies based on whether such knowledge is taken into account.



Actual IQA dialogues often exhibit "context-dependent" follow-up questions (FU Qs) containing anaphoric devices, like Q2 below. Such questions are potentially difficult to process by means of standard QA techniques, and it is for these cases that we claim that predicting the FU question's focus (here, the entity "library card") will help a system find the correct answer (cf. Sec. 6 for empirical backup).

**Q1:** Can high-school students use the library?
**A1:** Yes, if they got a library card.
**Q2:** So, how do I get it?

Following (Stede and Schlangen, 2004), we refer to the type of IQA dialogues we are studying as "information-seeking chat", and conjecture that this kind of dialogue can be handled by means of a simple model of discourse structure. Our assumption is that in general the user engages in a coherent dialogue with the system. As proposed in (Ahrenberg et al., 1995), we model the dialogues in terms of pairs of initiatives (questions) and responses (answers), ignoring other intentional acts.

The approach we adopt aims at answering the following questions: (a) In what way does information about the previous user questions and previous system answers help in predicting the next FU Q? (b) Does the performance of an IQA system improve if it has structure/history-based information? (c) Which is the role that each part of this information plays for determining the correct answer to a FU Q?

This paper is structured as follows. Section 2 gives an overview of some theories of focus used in dialogue and IQA. Section 3 then gives a detailed account of our theory, explaining *what* a question can focus on, and what patterns of focus change we expect a FU Q will trigger. Hence, this first



*Coling 2008: Proceedings of the workshop on Knowledge and Reasoning for Answering Questions,* pages 25–32
Manchester, August 2008

part answers our question (a) above. We then move to more applied issues in Sec. 4, where we show how questions and answers were annotated with focus information. The next Section 5 explains the Q→A algorithm we use to test our theory so as to answer (b), while Section 6 covers the logistic regression models with which we learn optimal values for the algorithm from data, addressing question (c).

## 2 Coherence in IQA dialogues

In the area of Discourse processing, much work has been devoted to formulating rules that account for the coherence of dialogues. This coherence can often be defined in terms of *focus* and *focus shifts*. In the following, we adopt the definition from (Lecœuche et al., 1999): *focus* stands for the "set of all the things to which participants in a dialogue are attending to at a certain point in a dialogue".[1] In general, all theories of dialogue focus considered by Lecœuche *et al.* claim that the focus changes according to some specific and well defined patterns, following the rules proposed by the respective theory. The main difference between these theories lies in how these rules are formulated.

A major distinguishing feature of different focus theories has been the question whether they address global or local focus. While the latter explain coherence between consecutive sentences, the former are concerned with how larger parts of the dialogue can be coherent. We claim that in "information seeking dialogue" this distinction is moot, and the two kinds of foci collapse into one. Furthermore, our empirical investigation shows that it suffices to consider a rather short history of the dialogue, i.e. the previous user question and previous system answer, when looking for relations between previous dialogue and a FU Q.

Salient transitions between two consecutive questions are defined in (Chai and Jin, 2004) under the name of "informational transitions". The authors aim to describe how the topic within a dialogue evolves. They take "entities" and "activities" as the main possible focus of a dialogue. A FU Q can be used to ask (i) a similar question as the previous one but with different constraints or different participants (topic extension); (ii) a question concerning a different aspect of the same topic (topic exploration); (iii) a question concerning a related activity or a related entity (topic shift). We take this analysis as our starting point, extend it and propose an algorithm to automatically detect the kind of focus transition a user performs when asking a FU Q, and evaluate our extended theory with real dialogue data. Following (Bertomeu et al., 2006) we consider also the role of the system answer, and we analyze the thematic relations between the current question and previous question, and the current question and previous answer. Unlike (Bertomeu et al., 2006), we attempt to learn a model of naturally occurring thematic relations in relatively unconstrained IQA dialogues.

## 3 Preliminary Observations

### 3.1 What "things" do users focus on?

For all forthcoming examples of dialogues, questions and answers, we will base our discussion on an actual prototype IQA system we have been developing; this system is supposed to provide library-related information in a university library setting.

In the dialogues collected via an earlier Wizard-of-Oz (WoZ) experiment (Kirschner and Bernardi, 2007), we observed that users either seem to have some specific library-related *task* (action, e.g. "search") in mind that they want to ask the system about, or they want to retrieve information on some specific *entity* (e.g., "guided tour"). People tend to use FU Qs to "zoom into" (i.e., find out more about) either of the two. In line with this analysis, the focus of a FU Q might move from the task (action/verb) to the entities that are possible fillers of the verb's semantic argument slots.

Based on these simple observations, we propose a task/entity-based model for describing the focus of questions and answers in our IQA setting. Our theory of focus structure is related to the task-based theory of (Grosz, 1977). Tasks correspond to verbs, which are inherently connected to an argument structure defining the verb's semantic roles. By consulting lexical resources like PropBank (Palmer et al., 2005), we can use existing knowledge about possible semantic arguments of

[1] This definition is in line with how *focus* has been used in Computational Linguistics and Artificial Intelligence (hence, "AI focus"), originating in the work of Grosz and Sidner on discourse entity salience. We follow Lecœuche *et al.* in that focused elements could also be actions/tasks. We see the most salient focused element (corresponding to the "Backward-looking center" in Centering Theory) as the *topic* of the utterance. Accordingly, in the following we will use the terms *focus* and *topic* interchangeably; cf. (Vallduvi, 1990) for a survey of these rather overloaded terms.

the tasks we have identified.

We claim that actions/verbs form a suitable and robust basis for describing the (informational) meaning of utterances in IQA. Taking the main verb along with its semantic arguments to represent the core meaning of user questions seems to be a more feasible alternative to deep semantic approaches that still lack the robustness for dealing with unconstrained user input.

Further, we claim that analyzing user questions on the basis of their task/entity structure provides a useful level of abstraction and granularity for empirically studying informational transitions in IQA dialogues. We back up this claim in Section 6. Along the lines of (Kirschner and Bernardi, 2007), we aim for a precise definition of focus structure for IQA questions. Our approach is similar in spirit to (Chai and Jin, 2004), whereas we need to reduce the complexity of their discourse representation (i.e., their number of possible question "topics") so that we arrive at a representation of focus structure that lends itself to implementation in a practical IQA system.

### 3.2 How focus evolves in IQA

We try to formulate our original question, "Given a user question and a system response, what does a salient FU Q focus on?" more precisely. We want to know whether the FU Q initiates one of the following three transitions:[2]

**Topic zoom** asking about a different aspect of what was previously focused

1. asking about the same task and same argument, but different question type (e.g., search for books: Q: where, FU Q: how)
2. asking about the same entity (e.g., guided tour: Q: when, FU Q: where)
3. asking about the same task but different argument (e.g., Q: search for books, FU Q: search for journals)
4. asking about an entity introduced in the *previous system answer*

**Coherent shift** to a "related" (semantically, or: verb→its semantic argument) focus

1. from task to semantically related task
2. from task to related entity: entity is a semantic argument of the task
3. from entity to semantically related entity
4. from entity to related task: entity is a semantic argument of the task

**Shift** to an unrelated focus

From the analysis of our WoZ data we get certain intuitions about salient focus flow between some preceding dialogue and a FU Q. First of all, we learn that a dialogue context of just one previous user question and one previous system answer generally provides enough information to resolve context-dependent FU Qs. In the remainder of this section, we describe the other intuitions by proposing alternative ways of detecting the focus of a FU Q that follows a salient relation ("Topic zoom" or "Coherent shift"). Later in this paper we show how we implement these intuitions as features, and how we use a regression model to learn the importance of these features from data.

**Exploiting task/entity structure** Knowing which entities are possible semantic arguments of a library-related task can help in detecting the focused task. Even if the task is not expressed explicitly in the question, the fact that a number of participant entities *are* found in the question could help identify the task at hand.

**Exploiting (immediate) dialogue context: previous user question** It might prove useful to know the things that the immediately preceding user question focused on. If users tend to continue focusing on the same task, entity or question type, this focus information can help in "completing" context-dependent FU Qs where the focused things cannot be detected easily since they are not mentioned explicitly. This way of using dialogue context has been used in previous IQA systems, e.g., the Ritel system (van Schooten et al., forthcoming).

**Exploiting (immediate) dialogue context: previous system answer** Whereas the role of the system answer has been ignored in some previous accounts of FU Qs (e.g., (Chai and Jin, 2004) and even in the highly influential TREC task (Voorhees, 2004)), our data suggest that the system answer does play a role for predicting what a FU Q will focus on: it seems that the system answer can introduce entities that a salient FU Q will ask more information about. (van Schooten and op den Akker, 2005) and (Bertomeu et al., 2006) describe IQA systems that also consider the previous system answer.

[2] Comparing our points to (Chai and Jin, 2004), Topic zoom: 1. and 2. are cases of topic exploration, 3. of topic extension, and 4. is new. Coherent shift: 1. and 2. are cases of topic shift, and 3. and 4. are new.

**Exploiting task/entity structure combined with dialogue context** It might be useful to combine knowledge about the task/entity structure with knowledge about the previously focused task or entity. E.g., a previously focused task might make a "coherent shift" to a participant entity likely; likewise, a previously focused entity might enable a coherent shift to a task in which that entity could play a semantic role.

The questions to be addressed in the remainder of the paper now are the following. Does the performance of an IQA system improve if it has structure/history-based information as mentioned above? Which is the role that each part of this information plays for determining the correct answer to a FU Q?

## 4 Tagging focus on three levels

Following the discussion in Section 3.1, and having studied the user dialogues from our WoZ data, we propose to represent the (informational) meaning of a user question by identifying the task and/or entity that the question is about (*focuses on*). Besides task and entity, we have Question Type (QType) as a third level on which to describe a question's focus. The question type relates to what type of information the user asks about the focused task/entity, and equivalently describes the exact *type of answer* (e.g., why, when, how) that the user hopes to get about the focused task/entity. Thus, we can identify the focus of a question with the triple <Task, Entity, QType>.

We have been manually building a small domain-dependent lexical resource that in the following we will call "task/entity structure". We see it as a miniature version of the PropBank, restricted to the small number of verbs/tasks that we have identified to be relevant in our domain, but extended with some additional semantic argument slots if required. Most importantly, the argument slots have been assigned to possible filler entities, each of which can be described with a number of synonymous names.

**Tasks** By analyzing a previously acquired extensive list of answers to frequently-asked library-related questions, we identified a list of 11 tasks that library users might ask about (e.g. search, reserve, pick up, browse, read, borrow, etc.). Our underlying assumption is that the focus (as identified by the focus triple) of a question is identical to that of the corresponding answer. Thus, we assume the focus triple describing a user question also describes its correct answer. For example, in Table 1, A1 would share the same focus triple as Q1.

We think of the tasks as abstract descriptions of actions that users can perform in the library context. A user question *focuses on* a specific task if it either explicitly contains that verb (or a synonym), or implicitly refers to the same "action frame" that the verb instantiates.

**Entities** Starting from the information about semantic arguments of these verbs available in PropBank, and extending it when necessary for domain-specific use of the verbs, for each task we determined its argument slots. Again by inspecting our list of FAQ answers, we started assigning library-related entities to these argument slots, when we found that the answer focuses on both the task and the *semantic argument* entity. We found that many answers focus on some library-related entity without referring to any task. Thus, we explicitly provide for the possibility of a question/answer being about just an entity, e.g.: "What are the opening times?". A user question focuses on a specific entity if it refers to it explicitly or via some reference phenomenon (anaphora, ellipsis, etc.) linked to the dialogue history.

**Question Types** We compiled a list of question (or answer) types by inspecting our FAQ answers list, and thinking about the types of questions that could have given rise to these answers. We aimed for a compromise between potentially more fine-grained distinctions of question semantics, and better distinguishability of the resulting set of labels (for a human annotator or a computer program).

We defined each question type by providing a typical question template, e.g.: "where: where can I find $Entity?", "whatis: what is $Entity?", "yesno: can I $Task $Entity?', "howto: how do I $Task $Entity?". Note how some question types capture questions that focus on some task along with some participant entity, while others focus on just an entity. We also devised some question types for questions focusing on just a task, where we assume an *implicit* semantic argument which is not expressed, e.g., "how can I borrow?" (where in the specific context of our application we can imply a semantic argument like "item"). A question has a specific question type if it can be paraphrased with the corresponding question template. An answer

has a specific type if it is the correct answer to that question template.

### 4.1 A repository of annotated answers

From our original collection of answers to library FAQs, we have annotated around 200 with focus triples. The triples we selected include all potential answers to the FU Qs from the free FU Q elicitation experiment described in the next section. Some of the actual answers were annotated with more than one focus triple, e.g., often the answer corresponded to more than one question type. The total of 207 focus triples include all 11 tasks and 23 different question types (where the 4 most frequent types were the ones mentioned as examples above, accounting for just over 50% of all focus triples).

For instance, the answer: "You can restrict your query in the OPAC on individual Library locations. The search will then be restricted e.g. to the Library of Bressanone-Brixen or the library of the 'Museion'." is marked by: <Task: search, Entity: specific library location, QType: yesno>.

The algorithm we introduce in Section 5 uses this answer repository as the set $A$ of potential candidates from which it chooses the answer to a new user question. Again, we assume that if we can determine the correct focus triple of a user question, the answer from our collection that has been annotated with that same triple will correctly answer the question.

### 4.2 Annotated user questions

Having created an answer repository annotated with focus triples, we need user questions annotated on the same three levels, which we can then use for training and evaluating the Q→A algorithm that we introduce in Section 5. We acquired these data in two steps: 1. eliciting free FU Qs from subjects in a web-based experiment, 2. annotating the questions with focus triples.

**Dialogue Collection Experiment** We set up a web-based experiment to collect genuine FU Qs. We adopted the experimental setup proposed in (van Schooten and op den Akker, 2005)), in that we presented to our subjects short dialogues consisting of a first library-related question, and a corresponding correct answer, as exemplified by "Q1" and "A1" in Table 1.

We asked the subjects to provide a FU Q "Q2" such that it will help further serve their information need in the situation defined by the given previous question-answer exchange. In this way, we collected 88 FU Qs from 8 subjects and 11 contexts (first questions and answers).[3]

**Annotating the questions** We annotated these 88 FU Qs, along with the 11 first questions that were presented to the subjects, with focus triples. By (informally) analyzing the differences between different annotators' results, we continuously tried to disambiguate and improve the annotation instructions. As a result, we present a pre-compiled list of entities from which the annotator selects the one they consider to be in focus, and that of all possible candidates is the one least "implied" by the context. Table 1 shows one example annotation of one of the 11 first user questions and two of the 8 corresponding FU Qs.

## 5 A feature-based Q→A algorithm

We now present an algorithm for mapping a user question to a canned-text answer from our answer repository. The decision about which answer to select is based on a score that the algorithm assigns to each answer, which in turn depends on the values of the features we have introduced in the previous section. Thus, the purpose of the algorithm is to select the best answer focus triple from the repository, based on feature values. In this way, we can use the algorithm as a test bed for identifying features that are good indicators for a correct answer. Our goal is to evaluate the algorithm based on its accuracy in finding correct focus triples (which are the "keys" to the actual system answers) for user questions (see Section 5.2).

For each new user question $q$ that is entered, the algorithm iterates through all focus triples $a$ in the annotated answer repository $A$ (cf. Section 4.1). For each combination of $q$ and $a$, all 10 features $x_{1,q,a} \ldots x_{10,q,a}$ are evaluated. Each feature that evaluates to true ($\beta = 1$) or some positive value, contributes with this score $\beta$ towards the overall score of $a$. The algorithm then returns the highest-scoring answer $\hat{a}$.

$$\hat{a} = \arg\max_{a \in A}(\beta_1 x_{1,q,a} + \cdots + \beta_{10} x_{10,q,a})$$

[3] In the future, we plan to collect real FU Qs from users of our online IQA system, which will solve the potential problem of these questions being somewhat artificial due to the experimental setting. However, we still expect our current data to be highly relevant for studying what users would ask about next.

| ID | Q/A | Task | Entity | QType |
|---|---|---|---|---|
| Q1 | Can I get search results for a specific library location? | search | specific library location | yesno |
| A1 | You can restrict your query in the OPAC on individual Library locations. (...) | | | |
| Q2a | How can I do that? | search | specific library location | howto |
| Q2b | How long is my book reserved there if I want to get it? | reserve | my book | howlong |

Table 1: Example annotation of one first question and two corresponding FU Qs

### 5.1 Features

Based on the intuitions presented in Section 3.2, we now describe the 10 features $x_{1,q,a}, \ldots, x_{10,q,a}$ that our algorithm uses as predictors for answer correctness. All Task and Entity matching is done using string matching over word stems. QType matching uses regular expression matching with a set of simple regex patterns we devised for our QTypes.

- 3 surface-based features $x_{1,q,a}, \ldots, x_{3,q,a}$: whether $\{\text{Task}_a, \text{Entity}_a, \text{QType}_a\}$ are matched in $q$. Entity feature returns the length in tokens of the matched entity.
- 1 task/entity structure-based feature $x_{4,q,a}$ : how many of the participant entities of $\text{Task}_a$ (as encoded in our task/entity structure) are matched in $q$.
- 4 focus continuity features $x_{5,q,a}, \ldots, x_{8,q,a}$: whether $\{\text{Task}_a, \text{Entity}_a, \text{QType}_a\}$ are continued in $q$, wrt. previous dialogue as follows:[4]
  - Task, Entity, QType continuity wrt. previous user question.
  - Entity continuity wrt. previous system answer.
- 2 task/entity structure + focus continuity features $x_{9,q,a}, x_{10,q,a}$:
  - Focused Task of previous user question has $\text{Entity}_a$ as a participant.
  - $\text{Task}_a$ has focused Entity of previous question as a participant.

### 5.2 First Evaluation

Table 2 shows manually set feature scores $\beta_1, \ldots, \beta_{10}$ we used for a first evaluation of the algorithm; we chose these particular scores after inspecting our WoZ data. With these scores, we ran the Q→A algorithm on the annotated questions of annotator 1, who had provided a "gold standard" annotation for 78 of the 99 user questions (the remainder of the questions are omitted because the annotator did not know how to assign a focus triple to them). For 24 out of 78 questions, the algorithm found the exact focus triple (from a total of 207 focus triples in the answer repository), yielding an accuracy of 30.8%.

| $k$ | $x_{k,q,a}$ | $\text{range}(x_{k,q,a})$ | $\beta_k$ |
|---|---|---|---|
| 1 | qTypeMatch | 0,1 | 4 |
| 2 | taskMatch | 0,1 | 3 |
| 3 | lenEntityMatch | n | 2 |
| 4 | nEntitiesInTask | n | 1 |
| 5 | taskContinuity | 0,1 | 1 |
| 6 | entityContinuity | 0,1,2 | 1 |
| 7 | qTypeContinuity | 0,1 | 1 |
| 8 | entityInPrevAnsw | 0,1,2 | 2 |
| 9 | entityInPrevTask | 0,1 | 1 |
| 10 | prevEntityInTask | 0,1 | 1 |

Table 2: Manually set feature scores

## 6 Logistic Regression Model

To improve the accuracy of the Q→A algorithm and to learn about the importance of the single features for predicting whether an answer from $A$ is correct, we want to learn optimal scores $\beta_1, \ldots, \beta_{10}$ from data. We use a logistic regression model (cf. (Agresti, 2002)). Logistic regression models describe the relationship between some predictors (i.e., our features) and an outcome (answer correctness).

We use the logit $\beta$ coefficients $\beta_1, \ldots, \beta_k$ that the logistic regression model estimates (from training data, using maximum likelihood estimation)

[4] Both entity continuity features evaluate to '2' when exactly the same entity is used again, but to '1' when a synonym of the first entity is used.

| | Coeff. | 95% C.I. |
|---|---|---|
| lenEntityMatch | 6.76 | 5.26–8.26 |
| qTypeMatch | 2.54 | 2.02–3.06 |
| taskContinuity | 2.17 | 1.39–2.94 |
| entityInPrevAnsw | 1.78 | 1.06–2.49 |
| taskMatch | 1.37 | 0.80–1.94 |
| prevEntityInTask | -1.24 | -2.06– -0.43 |

Table 3: Model $M_2$: Magnitudes of significant effects

for the predictors as empirically motivated scores. In contrast to other supervised machine learning techniques, regression models yield human-readable coefficients that show the individual effect of each predictor on the outcome variable.

### 6.1 Generating Training data

We generate the training data for learning the logistic regression model from our annotated answer repository $A$ (Sec. 4.1) and annotated questions (Sec. 4.2) as follows. For each human-annotated question $q$ and each candidate answer focus triple from our repository ($a \in A$), we evaluate our features $x_{1,q,a}, \ldots, x_{10,q,a}$. If the focus triples of $q$ and $a$ are identical, we take the particular feature values as a training instance for a correct answer; if the focus triples differ, we have a training instance for a wrong answer.[5]

### 6.2 Results and interpretation

We fit model $M_1$ based on the annotation of annotator 2 using all 10 features.[6] We then fit a second model $M_2$, this time including only the 6 features that correspond to coefficients from model $M_1$ that are significantly different from zero. Table 3 shows the resulting logit $\beta$ coefficients with their 95% confidence intervals. Using these coefficients as new scores in our Q→A algorithm (and setting all non-significant coefficients' feature scores to 0), it finds the correct focus triple for 47 out of 78 test questions (as before, annotated by annotator 1); answer accuracy now reaches 60.3%.

We interpret the results in Table 3 as follows. All three surface-based features are significant predictors of a correct answer. The length of the matched entity contributes more than the other two; we attribute this to the fact that there are more cases where our simple implementations of qTypeMatch and taskMatch fail to detect the correct QType or task. While the task/entity structure-based nEntitiesInTask clearly misses to reach significance, the history-based features taskContinuity and entityInPrevAnsw are useful indicators for a correct answer. The first is evidence for "Topic zoom", with the FU Q asking about a different aspect of the previously focused task, while the second shows the influence of the previous answer in shaping the entity focus of the FU Q. From the two "task/entity structure + focus continuity" features, we find that if a FU Q focuses on a task that in our task/entity structure has an argument slot filled with the previously focused entity, it actually indicates a *false* answer; the implications of this finding will have to be explored in future work.

Finally, to pinpoint the important contributions of structure- and/or focus continuity features, we fit a new model $M_3$, this time including only the 3 (significant) surface-based features. Evaluating the resulting coefficients in the same way as above, we get only 24 out of 78 correct answer focus triples, an accuracy of 30.8%. This result supports our initial claim that an IQA system improves if it has a way of predicting the focus of a FU Q.

## 7 Conclusion

Our original hypothesis was that a) knowledge about the dialogue history, and b) lexical knowledge about semantic arguments could improve an IQA system's ability to answer FU Qs. We operationalized these notions by formulating a set of 10 features that evaluate whether a candidate answer is the correct one given a new (FU) user question. We then used regression modeling to investigate the usefulness of each individual feature by learning from annotated IQA dialogue data, showing that certain knowledge about the dialogue history (the previously focused task, and the entities mentioned in the previous system answer) and about semantic arguments are useful for distinguishing correct from wrong answers to a FU Q. Finally, we evaluated these results by showing how our Q→A mapping algorithm's answer accuracy improved by using the empirically learned scores for all statistically significant predictors/features. The features and the Q→A algorithm as a whole are based on a simple way to describe IQA questions

[5] Although in this way we get imbalanced data sets with $|A| - 1$ negative training instances for each positive one, we have not yet explored this issue further.

[6] We use annotator 2's data for training, and annotator 1's for testing throughout this paper.

in terms of focus triples. By showing how we have improved an actual system with learned feature scores, we demonstrated this representation's viability for implementation and for empirically studying informational transitions in IQA.

Although the IQA system used in our project is in several ways limited, our findings about how focus evolves in real IQA dialogues should scale up to any new or existing IQA system that allows users to ask context-dependent FU Qs in a type of "information seeking" paradigm. It would be interesting to see how this type of knowledge could be added to other IQA or dialogue systems in general.

We see several directions for future work. Regarding coherent focus transitions, we have to look into which transitions to different tasks/entities are more coherent than others, possibly based on semantic similarity. A major desideratum for showing the scaleability of our work is to explore the influence of the subjects on our data annotation. We are currently working on getting an objective inter-annotator agreement measure, using external annotators. Finally, we plan to collect a large corpus of IQA dialogues via a publicly accessible IQA system, and have these dialogues annotated. With more data, coming from genuinely interested users instead of experimental subjects, and having these data annotated by external annotators, we expect to have more power to find significant and generally valid patterns of how focus evolves in IQA dialogues.

## Acknowledgments

We thank Marco Baroni, Oliver Lemon, Massimo Poesio and Bonnie Webber for helpful discussions.

## References

Agresti, Alan. 2002. *Categorical Data Analysis*. Wiley-Interscience, New York.

Ahrenberg, L., N. Dahlbäck, and A. Jönsson. 1995. Coding schemes for studies of natural language dialogue. In *Working Notes from AAAI Spring Symposium*, Stanford.

Bertomeu, Núria, Hans Uszkoreit, Anette Frank, Hans-Ulrich Krieger, and Brigitte Jörg. 2006. Contextual phenomena and thematic relations in database QA dialogues: results from a wizard-of-oz experiment. In *Proc. of the Interactive Question Answering Workshop at HLT-NAACL 2006*, pages 1–8, New York, NY.

Chai, Joyce Y. and Rong Jin. 2004. Discourse structure for context question answering. In *Proc. of the HLT-NAACL 2004 Workshop on Pragmatics in Question Answering*, Boston, MA.

Grosz, Barbara Jean. 1977. *The representation and use of focus in dialogue understanding.* Ph.D. thesis, University of California, Berkeley.

Kirschner, Manuel and Raffaella Bernardi. 2007. An empirical view on iqa follow-up questions. In *Proc. of the 8th SIGdial Workshop on Discourse and Dialogue*, Antwerp, Belgium.

Lecœuche, Renaud, Chris Mellish, Catherine Barry, and Dave Robertson. 1999. User-system dialogues and the notion of focus. *Knowl. Eng. Rev.*, 13(4):381–408.

Palmer, Martha, Daniel Gildea, and Paul Kingsbury. 2005. The proposition bank: An annotated corpus of semantic roles. *Comput. Linguist.*, 31(1):71–106.

Stede, Manfred and David Schlangen. 2004. Information-seeking chat: Dialogue management by topic structure. In *Proc. of SemDial'04 (Catalog)*, Barcelona, Spain.

Vallduvi, Enric. 1990. *The Informational Component*. Ph.D. thesis, University of Pennsylvania, Philadelphia, PA.

van Schooten, Boris and Rieks op den Akker. 2005. Follow-up utterances in QA dialogue. *Traitement Automatique des Langues*, 46(3):181–206.

van Schooten, Boris, R. op den Akker, R. Rosset, O. Galibert, A. Max, and G. Illouz. forthcoming. Follow-up question handling in the IMIX and Ritel systems: A comparative study. *Journal of Natural Language Engineering*.

Voorhees, Ellen M. 2004. Overview of the TREC 2004 question answering track. In *Proc. of the 13th Text REtrieval Conference*.