ConvEx-DS: A dataset for conversational explanations in recommender systems

Diana C. Hernandez-Bocanegra, Jürgen Ziegler

University of Duisburg-Essen, Forsthausweg 2, 47057 Duisburg, Germany

Abstract

Conversational explanations are a novel and promising means to support users' understanding of the items proposed by a recommender system (RS). Providing details about items and the reasons why they are recommended in a conversational, language-based style allows users to question recommendations in a flexible, user-controlled manner, which may increase the perceived transparency of the system. However, little is known about the impact and implications of providing such explanations, using for example a conversational agent (CA). In particular, there is a lack of datasets that facilitate the implementation of dialog systems with explanatory purposes in RS. In this paper we validate the suitability of an intent model for explanations in the domain of hotels, collecting and annotating 1806 questions asked by study participants, and addressing the perceived helpfulness of the responses generated by an explainable RS using such intent model. Thus, we release an English dataset (ConvEx-DS), containing intent annotations of users' questions, which can be used to train intent classifiers, and to implement a dialog system with explanatory purpose in the domain of hotels.

Keywords

Recommender systems, explanations, conversational agent, user study, dataset

1. Introduction

Providing explanations of the rationale behind a recommendation can bring several benefits to recommender systems (RS), by increasing users' perception of transparency, effectiveness, and trust [1]. Although most explanations in RS are presented statically (i.e., using a fixed display in a single step), recent work has shown that providing interactive options for obtaining explanatory information can positively influence users' perception of RS [2]. Interactive options in explanations allow users to take control over the desired level of detail of the explanatory information, by means of a two-way communication, where users can indicate to the system the most relevant aspects on which explanations should focus. However, these possibilities are mostly limited to click-based options. Another kind of interactive approach to explanations is the conversational approach, in which users can express their questions in their own words. However, this has been, so far, much less explored.

While conversational approaches have already gained some attention in explainable artificial intelligence (XAI), and formal models of conversational explanations have been proposed to

IntRS'21: Joint Workshop on Interfaces and Human Decision Making for Recommender Systems, September 25, 2021, Virtual Event

diana.hernandez-bocanegra@uni-due.de (D. C. Hernandez-Bocanegra); juergen.ziegler@uni-due.de (J. Ziegler)

0000-0002-1773-2633 (D. C. Hernandez-Bocanegra); 0000-0001-9603-5272 (J. Ziegler)

this end [3, 4, 5], little is known about the type of explanation-related questions users would ask to a RS. Although several datasets exist that support the development of dialog and question-answering (QA) systems, these are generally focused on open domain-search (e.g. [6, 7]), or specific processes such as flight or hotel booking, without a focus on explanatory interaction as such. In particular, and to our knowledge, there are no publicly available datasets intended to support the development of an explanatory dialog system for RS, specifically, there are as yet no datasets for detecting the user's intent expressed in a question. We therefore collected 1806 questions that users asked to a RS and annotated them with intents according to an intent classification scheme developed for this purpose by [8]. The dataset contains questions about hotel recommendations and supports machine-learning based intent classification for explanatory conversational agents (CA).

Query intents are often characterized by means of intent classification schemes, which usually involve multiple dimensions (e.g. [9, 10]). This approach can facilitate the implementation of automatic intent detection procedures (i.e. those allowing to identify what information a user desires [10], so a proper answer can be generated), since detection can be solved by splitting the task into several less complex text classification tasks, one per each dimension. However, to implement text classifiers based on the intent model of [8], we still faced a challenge: even though some existing datasets could be useful for classifying values of the proposed dimensions (e.g. comparison or assessment), some relevant dimension values (or classes) are not annotated in those datasets, as discussed in depth in section 2.4.

We extend previous work of [8], who collected a small set of 82 human-generated questions about recommended items through a Wizard of Oz (WoOz) study. Their proposed model addresses two entities: *hotel* and *hotel feature*, and two main intent types: *system-related* intents (related to the algorithm, or the system input) and *domain-related* intents (related to hotels and their features). In turn, the *domain-related* type consists of the following dimensions, with several values each: *comparison* (a question could be comparative or not), *assessment* (whether question refers to facts, to a subjective evaluation or the reasons why an item is recommended), *detail* (whether the question refers to a single aspect or to the entire item), and *scope* (whether the question is about a single item, several items, or to the whole set of recommendations). The authors argue that an intent can be considered as a combination of values of these dimensions, and that reasonable answers can be generated when using such a scheme. For instance, the intent expressed by "Why are the rooms at Hotel X great?" would be: non-comparative (comparison) / why-recommended (assessment) / aspect (detail) / single (scope); and in consequence - assuming a review-based explanation method -, a possible answer could be: "because 96% of opinions about rooms are positive".

Given that the dimension-based intent model [8] was derived on a very small data set, it is still necessary to evaluate the validity of this proposal on a larger scale, i.e., the extent to which the model is able to accurately represent user intents given a larger set of questions. In particular, as an indirect measure of validity, we set out to evaluate perceived helpfulness of the responses generated by an RS implementing the intent model, under the assumption that if the system has adequately recognized the user's intent, it is able to generate a response that approximates the user's information need, and thus be considered, to some extent, helpful. Therefore, in this paper we aim to answer: **RQ1**: How valid is the dimension-based intent model proposed by [8], when taking into account a larger number of user-generated questions?

To this end, we collected a corpus of 1806 questions, and evaluated the perceived helpfulness by users of the answers generated by the system we implemented for this purpose. Additionally, we annotated the intent of the collected questions, using guidelines inspired by the intent model definition. Our aim was twofold: 1) to train classifiers with a view to future developments and further empirical validation of the conversational approach, and 2) to further validate whether the intent model could generalize to a larger scale. More specifically: **RQ2**: To what extent the collected questions could be consistently classified by human annotators?

To answer this question, we calculated inter-annotator agreement and assessed the pattern of questions where agreement was low, as well as particular observations that arose during the annotation process (detailed in section 4).

Finally, we consolidated the intent gold standard for each question, and validated the performance of intent detection procedures trained using the final annotated corpus. More specifically: **RQ3**: To what extent does the intent classification perform better when trained on our annotated dataset, compared to the auxiliary datasets we used during corpus collection?

To answer our research questions, we implemented an explanatory RS, which could interpret user queries and provide answers based on the underlying RS algorithm used ([11]), and text classifiers for the different dimensions, based on the state-of-the-art natural language processing (NLP) model BERT [12]. These classifiers were initially trained on *auxiliary* datasets that could be useful for detecting certain (but not all) dimension values (as detailed in section 3.1). We then conducted a user study aiming both to collect a large number of user queries, and to evaluate the perceived helpfulness by users of system generated answers. Details of system implementation and corpus collection procedure are addressed in section 3. Finally, the contributions of this paper can be summarized as follows:

- We release ConvEx-DS [1] (**Conv**ersational **Ex**planations **D**ata**S**et), consisting of 1806 user questions with explanatory purpose in the domain of hotels, with question intent annotations, which can facilitate the development of explanatory dialog systems in RS.
- We implemented a RS that generates answers to these questions, and tested the user-perceived helpfulness of system generated answers.

2. Related work

2.1. Explanations in RS

Providing explanations for recommended items can serve different purposes. Explanations may enable users to to understand the suitability of the recommendations, to understand why an item was recommended, or they may assist users in their decision making. Among the most popular approaches are the methods based on collaborative filtering (e.g. "Your neighbors' ratings for this movie" [13]), as well as content-based methods that allow feature-based explanations, showing users how relevant item features match their preferences (e.g. [14]). On the other hand, review-based explanations usually show summaries of the positive and negative opinions about items (e.g. [15, 16, 17]). Our work is related to the latter approach, and our implemented

[1] ConvEx-DS can be downloaded at https://github.com/intsys-ude/Datasets/tree/main/ConvEx-DS

system uses the explanatory RS method proposed by [11], to generate both recommendations and explanations, based on ratings and customers' opinions.

2.2. Interactive and conversational explanations

In contrast to static approaches to explanations (which are dominant in RS and XAI overall [18]), interactive approaches seek to provide users with greater control over the explanatory components [2, 19], so that a better understanding of the reasons behind the recommendations can be achieved.

Moreover, conversational approaches to explanations take into account the social aspect of this process [20], where "someone explains something to someone" [21], through an exchange of questions and answers between the user and the system, as would occur in a human conversation. To this end, formal specifications and dialogue models of explanation (e.g. [3, 22, 5]) have been proposed as a theoretical basis for designing conversational explanations in intelligent systems. However, due to lack of sufficient empirical evaluation of such approaches [20, 4], it is still unclear how conversational explanatory interfaces should be conceived and designed in RS.

Recently, and inspired by dialog models of explanation [23], [8] proposed a dialog management policy and an user intent model, to implement a CA for explanatory purposes in a hotel RS. Our work builds on this model, and we extended this work by evaluating the intent model validity on a larger scale. While the prior work was based on the Wizard of Oz (WoOz) method for collecting user questions followed by explanations given by the experimenter, resulting in a set of 82 questions, in the present work we implemented a system to automatically generate answers, which allowed us to collect a larger number of questions (1806).

2.3. Intent detection and slot filling

We developed an RS system able to reply automatically to users' questions as part of an explanatory conversation. To this end, we set our focus on the natural language processing (NLP) tasks: intent detection and slot filling, key tasks for the development of dialog systems. Intent detection seeks to interpret the user' information need expressed through a query, while slot filling aims to detect which entities - and also features of an entity - the query refers to. The idea behind the intent concept is that user utterances within a dialogue can be framed within a finite and more limited set of possible dialogue acts. The most common approach for intent representation, in the open-search domain, is intent classification [10], that is, a query can be categorized according to a classification scheme, consisting of dimensions or categories, and their possible values, as in [9, 10]. The intent model by [8], on which we base our work, falls within this type of representation.

A large body of previous work has addressed the task of intent detection, both for open search domains (see [24, 10]) and task-oriented dialog systems, for processes such as flight booking, music search or e-banking, e.g. [25, 26, 27]. Methods proposed to solve these tasks range from conventional text classification methods, to more complex neural approaches, based on recurrent neural networks, attention-based mechanisms and transfer learning, to solve the intent detection and slot-filling tasks, both jointly and independently, and to extend the solutions to new domains. Since an in-depth comparison of the different approaches is beyond the scope of this paper, we refer readers to the survey on this matter by [28].

In particular, our work is related to the text classification approach, which leverages the representation of possible intents according to a classification scheme. According to this approach, the difficult task of intent detection can be divided into smaller text classification tasks, to detect the class that best represents a sentence according to each dimension. For this purpose, we implemented text classifiers using the state of art natural language processing model BERT [12]. As for the slot-filling task, and in line with [29], we solve it as a named entity recognition (NER) task. In our case, the entities to be recognized correspond to the names of the hotels about which the questions are asked. For this purpose, we use the NLTK toolkit [30].

2.4. Datasets for Intent detection

Benchmarking of intent-detection tasks is usually based on prominent datasets like ATIS [25] (Airline Travel Information System, containing queries related to flight searching), the MIT corpus [31] (queries to find movie information, or booking a restaurant), or the SNIPS dataset [26] (to develop digital assistants, involving tasks as asking for weather, or playing songs). To our knowledge, no public dataset has been published to support the development of dialog systems with explanatory purpose in RS. However, we investigated existing data sets that could contribute to classifying values along the different dimensions of the intent model.

Dimension comparison: Work by Jindal and Liu [32] addressed the identification of comparative sentences as a classification problem. The authors released a dataset with comparative and non-comparative sentences extracted from user reviews on electronic products, from forums involving comparison between brands or products, and from news articles on random topics. On the other hand, Panchenko et al. [33] released a dataset for comparative argument mining, involving 3 classes (better, worse or none), in domains like computer science, food or electronics. This dataset allows automatic detection of comparative sentences where entities to compare are explicitly mentioned (e.g. “Python is better suited for data analysis than MATLAB”), while superlative sentences like “which is the best option?” are not considered as comparative. The above was problematic for our purposes, since most comparative questions in the WoOz [8] set are precisely superlative. Consequently, we opted to use Jindal and Liu (see section 3.1).

Dimension assessment: Bjerva et al. [34] released SubjQA, a dataset for several domains (including hotels), which can be used to detect subjectivity of questions, in QA tasks. This dataset includes annotations of subjective and non-subjective classes, which can be leveraged to classify *evaluation* and *factoid* questions, according to the intent model by [8]). This dataset does not involve questions of the type *why-recommended*.

Dimension detail: While most aspect-based approaches involve the detection of an aspect or specific feature addressed in a sentence (e.g., room, facilities), as in [35, 2], detecting the absence of aspect is not usually addressed. Consequently, to our knowledge, no dataset involves annotation of sentences that addressed the quality of an overall item (e.g., "how good is Hotel x?"), in contrast to aspect-based sentences. Therefore, we used sentences collected in the WoOz study and the classification “detail”, as described in Section 3.1.

Dimension scope: To our knowledge, there is no dataset for the detection of the scope dimension. However, the values under this dimension can be inferred from entity detection (particularly hotels), for which NER can be used.

3. Corpus collection

Aiming to validate the intent model proposed by [8], we implemented and tested a conversational RS, consisting of a natural language understanding (NLU) module, which interprets questions with explanatory purpose written by users, and a module to generate answers consistent with the review-based recommendation method on which the RS is based. The development of our system and the corresponding user study involved a process consisting of several iterations. After every iteration, participants were asked to interact with the latest version of our system, so results of each iteration were used to improve the system to be tested in the next iteration. This was done in order to improve the performance of the classifiers, and to include new methods of response generation, for example to respond to intents that were not initially implemented, given their low frequency among all the questions asked by users. In addition to collecting participants' questions, we also captured their perception of the helpfulness of the answers generated by the system. Details of the methods implemented and the user study below.

3.1. Intent detection: methods and datasets

We divided the intent detection task into a set of three classification tasks (one for each of the dimensions: *comparison*, *assessment* and *detail*), and one NER task (for the detection of "hotel" entities, which allowed us to infer the *scope* dimension). Thus, the final detected intent corresponds to the combination of the values detected of each dimension. Thus, for example, the intent of the sentence "how good is the service at Hotel X" should be detected as: non-comparative (comparison) / evaluation (assessment) / aspect (detail) / single (scope).

BERT-based Text classifiers: We trained BERT classifiers [12], one for each dimension (comparison, assessment and detail), using a 12-layer model (*BertForSequenceClassification*, bert-base-uncased), batch size 32, and Adam optimizer (learning rate = 2e-5, epsilon = 1e-8). Classifiers for comparison and assessment converged after 4 epochs, while for detail 5 epochs were needed. Datasets were split randomly into training (80%) and test (20%) during the training phase. In order to avoid overfitting, the most represented class was downsampled (randomly) to approximate the size of the less represented class, which was slightly upsampled (randomly) to fit round numbers like 1000 and 500. In the case of the detail dimension, due to the small size of the auxiliary dataset, both classes were increased to 100 instances each (described below). Datasets (original and balanced) sizes are reported in table 1.

Dimension comparison: To train the classifier, we used the dataset by Jindal and Liu [32], which involves 5 classes (non-equal gradable, equative, superlative, non-gradable and non-comparative), all except the last one correspond to a detailed level of granularity for the sentences considered as comparative, which we believe is not necessary for our purposes. Thus, we grouped the sentences of the comparative classes (non-equal gradable, equative, superlative, and non-gradable), under a single *comparative* class.

Dimension assessment: We used the dataset by Bjerva et al. [34], specifically the one corresponding to the domain of hotels. Dataset includes an annotation whether the sentence is subjective or not, which we used to classify questions as evaluative and factoid, respectively. As this dataset does not involve the class *why-recommended*, we included a handcrafted validation, so subjective questions including the word "why" were regarded as such.

Table 1
Size and distribution of datasets used to train initial classifiers, implementation used during corpus collection phase.

	Dataset size	
Comparison [32]	**Original**	**Balanced**
Comparative	853	1000
Non-Comparative	7200	1000
Subjectivity [34]	**Original**	**Balanced**
Subjective	2706	500
Non-subjective	488	500
Detail [8]	**Original**	**Augmented, balanced**
Aspect	58	100
Overall	22	100

Dimension detail: Here, we leveraged questions collected by [8] in their WoOz study. As the size was extremely low, we used an augmentation technique, to generate synthetically new sentences from those in the WoOz dataset, altering some words, such as hotel names or aspects. Additionally, after initial iterations of the user study, we manually classified the collected questions written by participants as *aspect* or *overall*, added new sentences from the less represented class (i.e. the *overall*) to the dataset and retrained the *detail* classifier, so the risk of overfitting due to imbalanced classes and augmentation techniques could be decreased in the next iteration.

Dimension scope and entity 'hotels': First, we identify the entities (hotels) mentioned in the sentence. For this NER task, we used procedures from the library NLTK [30] to identify the entities (particularly the tokenizer and the part of the sentence (POS) tag methods). Then, we inferred the scope value depending on the number of entities recognized: *single* for one entity, *tuple* for more than two, and *indefinite* if no entity was found. An special case are the anaphoras regarding the entity [36], e.g. the sentence "how is the service at *this* hotel?" might refer to a previous hotel mentioned in the previous question or its answer. Usually, these situations are handled by the dialog system, which is in charge of keeping track of context. As a solution, when no entity was detected, but the sentence included a determiner such as 'this', 'these', 'those', 'its', 'their', etc, and if an entity was recognized or included in the previous question or its answer, the sentence was marked as *single* or *tuple*.

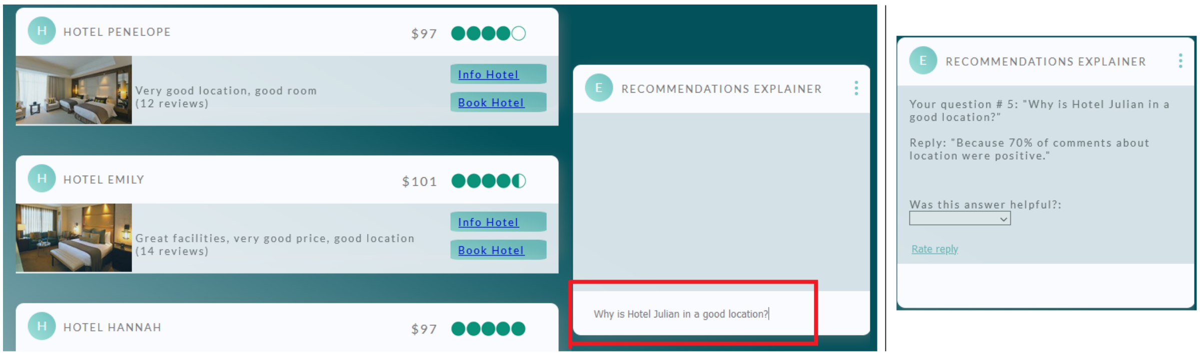
Entity 'hotel feature': We used the aspect-based detection methods implemented by [2], which use BERT classifiers and the ArguAna dataset [35], to detect aspect and hotel features addressed in user questions.

3.2. Explainable RS

We used the review-based RS developed by [2], which implements the matrix factorization model proposed by [11], in combination with sentiment-based aspect detection methods, using the state of art NLP model BERT [12], in order to provide aspect based arguments. We also use the personalization mechanism described by [2], which uses the aspects reported as preferred by participants in the study survey, to generate personalized recommendations.

Answer generation module: We implemented a module to generate replies based on the intent detected, and based on the type of argumentative responses proposed by [2]. According to this proposal, *factoid* questions could be replied with Y/N or a value (e.g. check in times) based on metadata. As for *evaluation* or *why-recommended* questions, replies were based on the aggregation of positive or negative opinions regarding an aspect (if question was *aspect* based), or the most important aspects for the participant (in case question was *overall*). This aggregation of opinions was calculated based on the hotels the question was about. If the question was comparative, the system calculated which hotel was better among a tuple, or the best in general (scope *indefinite*), based on the aggregation of the opinions. These are some examples of the type of responses generated by the system: Q: "Why does Hotel Hannah have the highest rating?", A: "Because of the positive comments reported regarding the aspects that matter most to you: 86% about location, and 85% about price."; Q: "Which hotel is best, Hotel Lily, Hotel Amelia or Hotel Hannah?", A: "Hotel Lily has better comments on the aspects that are most important to you (location, facilities, staff). However, Hotel Amelia has better comments about room, price."; Q: "Hotel Amelia is described as having a great room, what makes it great?, A: "Comments about rooms are mostly positive (90%).".

3.3. User study



Very good location, good room
(12 reviews)

Great facilities, very good price, good location
(14 reviews)

Why is Hotel Julian in a good location?

Your question # 5: "Why is Hotel Julian in a good location?"

Reply: "Because 70% of comments about location were positive."

Was this answer helpful?:

Figure 1: Interface system used for corpus collection. Left, list of recommendations and box to write questions (highlighted in red). Right, system shows answer and requests users to rate helpfulness

Participants: We recruited 298 participants (209 female, mean age 30.42 and range between 18 and 63) through the crowdsourcing platform Prolific. We restricted the task to workers in the U.S and the U.K., with an approval rate greater than 98%. Participants were rewarded with 1.5 plus a bonus up to 0.30 depending on the quality of their response to the question "Why did you choose this hotel?" set at the end of the survey, aiming to achieve a more motivated hotel choice by participants, and encouraging effective interaction with the system. Time devoted to the task (in minutes): M=7.31, SD= 4.97. Questions asked per participant: M=5.99, SD=2.58.

We applied a quality check to select participants with quality survey responses (we included attention checks in the survey, e.g. "This is an attention check. Please click here the option 'yes'"). Users were told in the instructions that at least 5 questions were required as a prerequisite for payment, as well as correct answers for the attention checks (2). We discarded participants with at least 1 failed attention check, or no effective interaction with the system, i.e. if users did not ask questions to the system. Thus, the responses of 41 of the 339 initial participants were discarded and not paid (final sample: 298 subjects).

Procedure: Users were asked to report a list of their 5 most important aspects when looking for a hotel, sorted by importance. Then we presented participants with instructions: 1. They would be presented with a list of 10 hotels with the results of a hypothetical search for hotels already performed using a RS (i.e., no filters were offered to search for hotels). 2. They could consult general hotel information (photos, reviews, etc., by clicking on "Info Hotel"), but indicated that we were more interested in knowing their questions about the reasons why these hotels were recommended, stating that "The aim of the system is to provide explanations based on your questions" (we aimed here to prevent the user from asking questions about other processes, such as booking assistance). 3. They should write each question (in their own words) at the bottom of the explanation box (highlighted in red in the example), and click enter to send (see Fig. 1 left). 4. Next, the system would present the answer to their question, and a drop-down list to evaluate how helpful you think the answer was (with values from "Strongly disagree" to "Strongly Agree"). They had to choose a value from the list and click on the "Rate Reply" link, continue with your next question, and repeat until they complete at least 5 questions (Fig. 1 right). 6. Once they finished, they had to indicate which hotel they would finally choose by clicking the button "book hotel". 7. Back on the survey, they had to describe why they chose that hotel, we stated that a bonus would be paid depending on the quality of this response. A reminder of these instructions was included in the app, so it would be easier for users to remember them. After instructions and before the task, we presented a cover story, to establish a common starting point in terms of travel motivation (a holiday trip). The question used to rate the usefulness of the system's answers was: *"Was this answer helpful?"*, and reply was measured with a 1-5 Likert-scale (1: Strongly disagree, 5: Strongly agree).

4. Corpus annotation

4.1. Intent type annotation

First, sentences where classified according to the classes: domain-related intents (regarding hotels and their features), and system-related intents (regarding the algorithm, the system Input, or system functionalities). [8] reported that domain questions clearly outnumbered system questions, so the research team members annotated this class instead of crowdsourcing workers (98.3% agreement), as the low number of system questions could lead to the category being ignored in the crowdsourcing setup. Disagreements were resolved in joint meetings.

4.2. Dimension-based annotation

Only domain-related sentences were used for the dimension-based annotation. We collected annotations for comparison, assessment and detail as independent tasks. The dimension scope was not annotated under the proposed procedure (is not a classification task but a NER task).

Annotators and crowdsourcing setup: Every sentence was annotated by 3 annotators: one belonging to the research team, and the other two crowdsourced on the Prolific platform. We divided the set of questions into 19 blocks of 100 sentences each, and every block had to be annotated for each dimension separately, to mitigate the fatigue associated with a longer list of questions, which could affect the participant's performance. Each block included 4 attention checks (e.g. "This is an attention check. Please click here the option 'comparative'"). Participants were warned that failing this check or not completing the list of 100 questions would lead to rejection and non-payment of the task. We also included questions from the examples provided in the guidelines within the blocks, for a subsequent attention check (failing this check led to rejection of the block for the agreement and final gold standard).

The research team annotator annotated all blocks for the three dimensions, while different crowdsourcing workers could annotate different blocks for different dimensions. Same annotator did not annotate the same block for the same dimension more than once. This way we ensured that each sentence was annotated by 3 different people, for each dimension.

Procedure: Once participants took the task in Prolific, they had to read the instructions of the task (annotation guidelines), and then open the annotation application (where annotation guidelines remained visible). Once the end (100 questions) was reached, the user was prompted to return to the main survey, and to report observations or difficulties.

Annotation application: We developed a simple annotation application, in which annotators could select the class to which a question belongs, according to each dimension. The user interface consisted of a single page, showing: at the top, a reminder of the guidelines for annotation; at the bottom, the consecutive number of the question (so that the user could note its progress, e.g. 2 out of 100), the question, a checkbox to indicate the class to which the sentence belonged, and a "Next question" button.

Participants, and selection of valid submissions: 92 participants performed the annotation task using the platform Prolific. We restricted the task to workers in the U.S and the U.K., approval rate greater than 98%. Participants who annotated comparison blocks were rewarded with 1.25, assessment blocks with 1.50, and detail blocks with 1.25. Differences in payment were due to the different devoted times in minutes for each dimension (comparison: M=9.85, SD=3.53, assessment: M=13.24, SD=5.86, detail: M=9.40, SD=2.69). Participants who failed the attention checks, or those who did not complete the task, were rejected and not paid (19 participants in total). None of the questions submitted by these participants were used in the final calculation of the gold standard, nor for the agreement score. As part of a subsequent quality check, we discarded participants and their submitted answers, if they failed to correctly classify questions that also appeared as examples in the instructions, although their submissions were paid (16 participants). No further criteria were used to discard blocks of user responses, as we were not to establish correct or incorrect answers, but to establish whether the elaborated guides were understood in a similar way by different users, and whether the classes established by the intent model fit the questions in the corpus. A final set of annotations by 57 Prolific workers was used for the calculation of Inter-rater reliability, and the deduction of gold standard.

Classifiers trained on ConvEx-DS: Bert model, batch size, Adam optimizer parameters, and splitting as reported in section 3.1. To avoid overfitting, the most represented class was downsampled (randomly), to approximate the size of the less represented class. Classifiers of comparison and assessment dimensions converged after 4 epochs, of dimension detail after 5.

5. Results

5.1. Helpfulness of system answers

Taking into account iterations 4 to 6 (most refined versions of the system) the system was able to generate an answer in 80.58% of the cases, and to partially recognize the intent or entities in 7.34% of the cases (thus asking the user to rephrase or indicate further information). Among the main reasons why the questions were not replied we found: complexity of the question or not information available to reply (31%), text that could be improved when replying factoid questions (23%), wrong intent classification (11%), and system errors (11%).

Figure 2 (middle) shows the perception of answers' helpfulness, according to ratings granted by study participants, across all iterations (M=3.58, SD=1.34). When taking into account only the last two iterations (which account for 63.85% of sentences collected, and involve the most refined versions of the system), we observed a greater perceived helpfulness (to M=3.70, SD=1.30). We considered as "non-helpful" responses those that were marked with the values "Strongly Disagree" and "Disagree" when participants were asked "Was the answer helpful?". We analyzed the responses given to those questions by the system and found that in 34% of cases, replies provided actually make sense, i.e. seemed a reasonable answer to the asked question. Among the reasons that caused the responses to be rated as non-helpful, we found: 30% due to misclassified intents or entities, 14% to system errors, 9% text that could be improved when replying factoid questions, 5% due to complexity of the question or not information to reply it, and 5% to specific aspects not addressed by the solution.

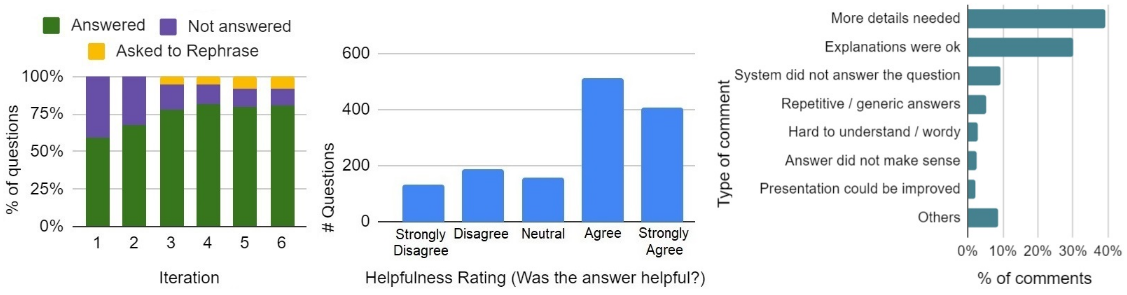


Figure 2: Left: Distribution of replied questions, across iterations. Middle: Histogram of helpfulness rating granted by users to answers generated by the system (all iterations). Right: Types of comments by participants during corpus collection, in regard to system answers.

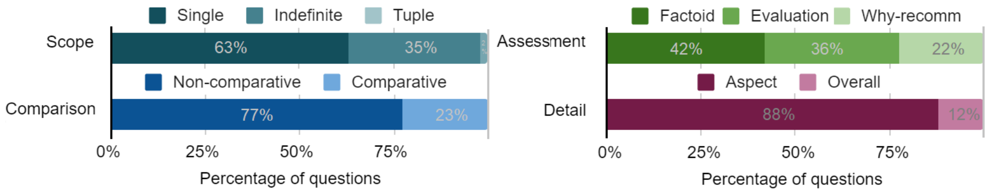


Figure 3: Distribution of questions in ConvEx-DS (domain-related intents).

Table 2
Inter-rater reliability of ConvEx-DS. *Fleiss' kappa* refers to each dimension, and the *% of full agreement* to each class, i.e. percentage of questions in which all annotators agreed on assigning that class).

Dimension	Fleiss' kappa	Class	% of full agreement
Comparison	0.72	Comparative	77.28%
		Non-comparative	86.86%
Assessment	0.65	Factoid	73.99%
		Evaluation	58.56%
		Why-recommended	66.42%
Detail	0.75	Aspect	95.73%
		Overall	66.82%

5.2. Annotation statistics

A total of 1836 questions were collected during the corpus collection step. 30 of those questions were discarded (nonsense statements, or highly ungrammatical), for a final set of 1806 of annotated questions. Of these, only 24 were annotated as system-related questions. Length of questions: characters M=39.2, SD=15.67, words M=7.35, SD=2.86. We found a Fleiss' kappa of 0.72 for *comparison*, of 0.65 for *assessment* and of 0.75 for *detail*, indicating a "substantial agreement" [37], for all three dimensions. As for classes with lower percentages of questions with full agreement, we identified the following main causes: **Dimension assessment:** - *Why-recommended* questions rated as subjective, given that adjectives like 'good' or 'great' are included in sentences, e.g. "why is hotel hannah location great?". - Questions that should be replied with a fact, but include adjectives that indicate subjectivity, e.g. "does hotel emily have any bad reviews?", "are there good transport links?", "which hotel best fits my needs?". - Questions with adjectives as 'cheap', 'expensive', 'close', 'near', 'far', which can be answered with either subjective or factoid responses, e.g. "Which is the cheapest hotel?", "is there an airport near any of these hotels?". - Questions of the type "what is ... like", e.g. "What is the room quality like at Hotel Emily?" (this type of questions were actually not addressed in instructions).

Dimension detail: - Concepts that were regarded as hotel aspects, e.g. value (Which is best value for money), ratings (What is the highest rating for Hotel Levi), reviews (Which hotel has the most reviews?), stars (Which hotels are 5 stars?).

Finally, we have detected some questions that could hardly fit in the planned classes, e.g. "How do you define expensive? Do you compare against facilities and what is included in the price?", "The Evelyn has 17 reviews and a positive feedback but scores lower than others with less reviews. Why is this?". However, we found this number to be rather low (16 questions).

5.3. Intent detection performance

Dimensions comparison, assessment and detail: To verify the performance of classifiers, we have calculated F1, a measure of classification accuracy. We tested accuracy in 3 different steps: 1) performance of models trained on auxiliary datasets [32, 34, 8], used for the system used in corpus collection. 2) We tested these models using our newly obtained annotated data, ConvEx-DS. 3) We trained and tested new classifiers, based entirely on ConvEx-DS. We report F1 scores for each dimension (*comparison*, *assessment* and *detail*). We reported weighted average,

Table 3
F1-scores (weighted average) of classifiers of different dimensions, trained and tested on both auxiliary datasets and ConvEx-DS.

Dimension	Dataset	F1
Comparison	Jindal and Liu [32] [Training, Testing]	0.87
	Jindal and Liu [32] [Training], ConvEx-DS [Testing]	0.88
	ConvEx-DS [Training, Testing]	0.92
Assessment	Bjerva et al. [34] [Training, Testing]	0.93
	Bjerva et al. [34] [Training], ConvEx-DS (without why-recomm) [Testing]	0.60
	ConvEx-DS [Training, Testing]	0.91
Detail	WoOz augmented [Training, Testing]	0.98
	WoOz augmented [Training], ConvEx-DS [Testing]	0.90
	ConvEx-DS [Training, Testing]	0.92

to take into account the contribution of each class, which in (2) is particularly unbalanced (no downsampling of the test set was done, since balanced data was pertinent only for training).

We detected that the classifier trained on Bjerva et al. [34], performed particularly poorly when tested with our annotated data (ConvEx-DS). Here, we detected that 32% of questions under "evaluation" class in ConvEx-DS but classified as "non-subjective" correspond to questions regarding indefinite or more than two hotels (e.g. "which hotel has the best facilities?"), 18% corresponded to adjectives like "close", "far", "expensive", and 14% to questions of the form "what is the food like?". As of factoid questions in ConvEx-DS classified as subjective, we found 33% of questions involving indefinite or more than two hotels, and 32% regarded questions of the form "does the hotel have...". In section 6 we discussed these findings in depth.

Dimension scope: Entities (hotels) addressed in sentences were detected using the NLTK library. In order to check the accuracy of the method, 2 members of the research team have checked the inferred entity for the collected corpus, and found that in 5.38% of cases, the inferences were wrong. Most of these cases corresponded to cities, or facilities recognized as entities, a drawback detected in early stages of corpus collection, thus additional validations were added to the procedure, so that these cases would not occur in future iterations.

6. Discussion

To date, creating dialog systems able to answer all possible users' questions remains unrealistic [38]. Nevertheless, we found that in the later iterations of our user study, our implemented system was able to answer a wide number of questions, or to ask users to rephrase or better specify their explanation need. However, since the ability to answer the questions is not a sufficient condition for concluding that a model of intent is valid, we set out to validate how helpful the answers were perceived by users, as an indirect measure of model validity, assuming that a correct intent detection would lead - to a certain extent - to the responses being perceived as helpful. In this respect, we found that system answers were perceived as predominantly helpful, thus answering **RQ1** positively. On the other hand, ratings of non-helpfulness did not necessarily imply that the queries did not match the detected intent. In fact, we found that almost one-third of responses rated as non-helpful fitted the question (i.e. made sense). After a review of participants' feedback on the system answers, we found that, although many users found them helpful or "ok," the main criticism was that some of the answers lacked sufficient detail. For instance, it was not enough to simply answer yes / no to factoid questions, but further details about the inquired feature were expected. As for the *evaluation* or *why-recommended*, participants reported that the percentages of positive and negative opinions were fine, but some also demanded examples of such opinions. The above is consistent with findings reported by [2], who found that perception of explanation sufficiency was greater when options were offered to obtain excerpts from customer reviews, and that the need for more detailed explanations may depend on individual characteristics, for instance, decision-making style: users with a predominant rational decision making style have a tendency to thoroughly explore information when making decisions [39].

In consequence, although the intent model seems appropriate to generate an initial or first level response, a dialog system implementation must go beyond this initial response, offering options to drill down into the details. Similarly, criticized aspects, such as repetitive or too generic answers, could also be mitigated with such a solution, since providing excerpts of customer reviews as answers would allow a balance between system-generated and customer-generated statements. An alternative in this respect is to provide natural language explanations based on customer reviews, using abstractive summarization techniques as in [40]. However, as reported [41], users seem to prefer explanations that include numerical anchors (e.g. percentages) in comparison to only based text summaries, since percentages may convey more compelling information, while summaries may be perceived as too imprecise.

In line with [8], we also found that questions related with *system-related* intents were clearly outnumbered by *domain-related* intents, showing that in the explanatory context of hotels RS, users usually do not formulate questions explicitly addressing the system or its algorithm. We believe that users are highly less interested in such details, due to the nature of the domain addressed. Hotels are experience goods (those which cannot be fully known until purchase [42]), for which an evaluation process is characterized by a greater reliance on word-of-mouth [42, 43], which may lead users to grant much more attention to item features and customers' opinions about it, rather than on the details of the algorithm or how their own profile is inferred.

In regard to the annotation task, we found a substantial agreement in all the annotated dimensions, as well as a very encouraging accuracy measure, when classifiers were trained on the ConvEx-DS, which leads us to conclude, that under the intent model and annotation guidelines based on [8], the questions could be, to a substantial extent, unequivocally classified, thus replying to our **RQ2**. We note, however, the challenge of addressing the dimension assessment. In this regard, we found that the main difficulty was to classify correctly questions that could be regarded as evaluation, given their subjective nature (including expressions like "how close/far"), but for which a factual-based response could be given (e.g. "100 meters from downtown"), a similar concern raised by [34]: "a subjective question may or may not be associated with a subjective answer". Additionally, questions like "why is hotel X good?" were often classified as *evaluation*, given their subjective nature (adjective "good" as an indicator of subjectivity), so they were regarded as similar to their evaluation counterpart ("how good is hotel X?"). However, we believe that the distinction "why good" should be kept separate from "how good", since in the former, the user challenges arguments already provided by the system (a recommendation, or its explanations), while in the later this is not necessarily the case.

As for **RQ3**, we found that intent classifiers perform better when trained on ConvEx-DS, compared to classifiers trained on the auxiliary datasets, but tested on ConvEx-DS. Here, the most striking case concerns the dataset for the detection of subjective questions (SubjQA) by Bjerva et al. [34]. The above in no way suggests anything problematic in the SubjQA itself, only that in comparison to ConvEx-DS (dimension "evaluation"), the two datasets measure rather different concepts. SubjQA addresses the subjectivity of the question asked, not whether the question involves an *evaluation* that might be subjective, as in ConvEx-DS. Thus, for example, "how is the food?" is classified as non-subjective under SubjQA, since it does not contain expressions indicating subjectivity. Thus, non-subjective under SubjQA does not necessarily imply factoid. In addition, classifiers trained in SubjQA do not work well with questions that involve some sort of comparison between multiple items, since the SubjQA only involves questions addressing single items, for which an answer could be found in a single review.

Limitations: Despite our motivating results, it is important to note the limitations imposed by the discussed approach. Addressing intent detection as a text classification problem, by means of an intent classification model, allows to provide answers that approximate the information need expressed by the user. However, the approach is insufficient when dealing with questions that are too specific, particularly in regard to factoid questions. Consequently, the development of a DS with explanatory purposes in RS should not only rely on the underlying RS algorithm, customer reviews or hotels metadata (as in our developed system), but should also integrate further sources of information, e.g. external location services, in order to provide very specific details, like surroundings, distances to places of interest or transport means, in case these are not found in customer reviews or metadata.

Also, our study setup for corpus collection was based on a question/answer sequence (with helpfulness rating in between), thus not necessarily resembling a fluid chatbot-style dialog, in which users might write utterances, such as greetings or thanks, expressions that could not be classified under the intent model. Therefore, we suggest the use of alternative mechanisms for the detection and treatment of such expressions.

7. Conclusions and future work

Based on our results, we conclude that the dimension-based intention model proposed by [8] is a valid approach to represent user queries in the context of explanatory RS. We also believe that ConvEx-DS can significantly contribute to the development of dialog systems that support conversational explanations in RS.

As future work, we plan to explore the users' perception of a RS, where further details and excerpts from customers reviews are provided during the explanatory conversation, aiming to increase the perceived helpfulness by users of the responses that our system is able to generate. Additionally, although questions in ConvEx-DS involve only one domain, we believe it can also be leveraged for the development of explanatory approaches in RS for other domains, especially those involving review-based recommendations. In this sense, we plan to explore recent NLP developments, particularly on transfer learning techniques, to obtain linguistic representations that can serve as a basis for similar domains, particularly those where customer reviews are also exploited, such as restaurants, movies and shopping.

Acknowledgments

This work was funded by the German Research Foundation (DFG) under grant No. GRK 2167, Research Training Group “User-Centred Social Media”.

References

[1] N. Tintarev, J. Masthoff, Explaining recommendations: Design and evaluation, in: Recommender Systems Handbook, Springer US, Boston, MA, 2015, p. 353–382.

[2] D. C. Hernandez-Bocanegra, J. Ziegler, Effects of interactivity and presentation on review-based explanations for recommendations, in: Human-Computer Interaction – INTERACT 2021, Springer International Publishing, 2021, pp. 597–618.

[3] D. Walton, The place of dialogue theory in logic, computer science and communication studies 123 (2000) 327–346.

[4] P. Madumal, T. Miller, L. Sonenberg, F. Vetere, A grounded interaction protocol for explainable artificial intelligence, in: Proc. of the 18th International Conference on Autonomous Agents and Multiagent Systems, AAMAS 2019, 2019, p. 1–9.

[5] A. Rago, O. Cocarascu, C. Bechlivanidis, F. Toni, Argumentation as a framework for interactive explanations for recommendations, in: Proceedings of the Seventeenth International Conference on Principles of Knowledge Representation and Reasoning, 2020, p. 805–815.

[6] E. Merdivan, D. Singh, S. Hanke, J. Kropf, A. Holzinger, M. Geist, Human annotated dialogues dataset for natural conversational agents, Appl. Sci 10 (2020) 1–16.

[7] A. Ritter, C. Cherry, W. B. Dolan, Data-driven response generation in social media, in: Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing, 2011, p. 583–593.

[8] D. C. Hernandez-Bocanegra, J. Ziegler, Conversational review-based explanations for recommender systems: Exploring users’ query behavior (in press), in: 3rd Conference on Conversational User Interfaces (CUI ’21), 2021.

[9] A. Broder, A taxonomy of web search, ACM SIGIR Forum 36 (2002) 3–10.

[10] S. Verberne, M. van der Heijden, M. Hinne, M. Sappelli, S. Koldijk, E. Hoenkamp, W. Kraaij, Reliability and validity of query intent assessments: Reliability and validity of query intent assessments, Journal of the American Society for Information Science and Technology 64 (2013) 2224–2237.

[11] Y. Zhang, G. Lai, M. Zhang, Y. Zhang, Y. Liu, S. Ma., Explicit factor models for explainable recommendation based on phrase-level sentiment analysis, in: Proceedings of the 37th international ACM SIGIR conference on Research and development in information retrieval, 2014, p. 83–92.

[12] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding (2019).

[13] J. L. Herlocker, J. A. Konstan, J. Riedl, Explaining collaborative filtering recommendations, in: Proceedings of the 2000 ACM conference on Computer supported cooperative work, ACM, 2000, p. 241–250.

[14] J. Vig, S. Sen, J. Riedl, Tagsplanations: explaining recommendations using tags, in:

Proceedings of the 14th international conference on Intelligent User Interfaces, ACM, 2009, p. 47–56.

[15] K. I. Muhammad, A. Lawlor, B. Smyth, A live-user study of opinionated explanations for recommender systems, in: Intelligent User Interfaces (IUI 16), volume 2, 2016, p. 256–260.

[16] N. Wang, H. Wang, Y. Jia, , Y. Yin, Explainable recommendation via multi-task learning in opinionated text data, in: Proceedings of the 41st International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 18, 2018, p. 165–174.

[17] D. C. Hernandez-Bocanegra, J. Ziegler, Explaining review-based recommendations: Effects of profile transparency, presentation style and user characteristics, Journal of Interactive Media 19 (2020) 181–200.

[18] A. Abdul, J. Vermeulen, D. Wang, B. Y. Lim, M. Kankanhalli, Trends and trajectories for explainable, accountable and intelligible systems: An hci research agenda, in: Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems - CHI 18, 2018, p. 1–18.

[19] K. Sokol, P. Flach, One explanation does not fit all: The promise of interactive explanations for machine learning transparency 34 (2020) 235–250.

[20] T. Miller, Explanation in artificial intelligence: Insights from the social sciences, Artificial Intelligence (2018).

[21] D. J. Hilton, Conversational processes and causal explanation 107 (1990) 65–81.

[22] A. Arioua, M. Croitoru, Formalizing explanatory dialogues, Scalable Uncertainty Management (2015) 282–297.

[23] D. Walton, A dialogue system specification for explanation 182 (2011) 349–374.

[24] J. Hu, G. Wang, F. L. J. tao Sun, Z. Chen, Understanding user's query intent with wikipedia, in: Proceedings of the 18th international conference on World wide web - WWW '09, 2009.

[25] C. T. Hemphill, J. J. Godfrey, G. R. Doddington, The atis spoken language systems pilot corpus, in: In Proceedings of the workshop on Speech and Natural Language - HLT '90, 1990, p. 96–101.

[26] A. Coucke, A. Saade, A. Ball, T. Bluche, A. Caulier, D. Leroy, C. Doumouro, Snips voice platform: an embedded spoken language understanding system for private-by-design voice interfaces, in: ArXiv, abs/1805.10190, 2018.

[27] I. Casanueva, T. Temčinas, D. Gerz, M. Henderson, I. Vulić, Efficient intent detection with dual sentence encoders, in: arXiv:2003.04807, 2020.

[28] S. Louvan, B. Magnini, Recent neural methods on slot filling and intent classification for task-oriented dialogue systems: A survey, in: Proceedings of the 28th International Conference on Computational Linguistics, 2020, p. 480–496.

[29] R. Grishman, B. Sundheim, Message understanding conference- 6: A brief history, in: Proceedings of the 28th International Conference on Computational Linguistics, 2020, p. 466–471.

[30] E. Loper, S. Bird, Natural language processing with python: analyzing text with the natural language toolkit. (2009).

[31] J. Liu, P. Pasupat, S. Cyphers, J. R. Glass, Asgard: A portable architecture for multilingual dialogue systems, in: In IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2013, 2013, p. 8386–8390.

[32] N. Jindal, B. Liu, Identifying comparative sentences in text documents, in: Proceedings of

the 29th annual international ACM SIGIR conference on Research and development in information retrieval - SIGIR 06, 2006, pp. 244–251.

[33] A. Panchenko, A. Bondarenkoy, M. Franzekz, M. Hageny, C. Biemann, Categorizing comparative sentences, in: In Proceedings of the the 6th Workshop on Argument Mining (ArgMining 2019), 2019.

[34] J. Bjerva, N. Bhutani, B. Golshan, W.-C. Tan, I. Augenstein, Subjqa: A dataset for subjectivity and review comprehension, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing EMNLP, 2020, p. 5480–5494.

[35] H. Wachsmuth, M. Trenkmann, B. Stein, G. Engels, T. Palakarska, A review corpus for argumentation analysis, in: 15th International Conference on Intelligent Text Processing and Computational Linguistics, 2014, p. 115–127.

[36] S. Quarteroni, S. Manandhar, Designing an interactive open-domain question answering system, Natural Language Engineering 15 (2008) 73–95.

[37] J. R. Landis, G. G. Koch, The measurement of observer agreement for categorical data, Biometrics 33 (1977) 159–174. Klagenfurt, Germany: SSOAR.

[38] R. J. Moore, R. Arar, Conversational ux design: An introduction, Studies in Conversational UX Design (2018) 1–16. Springer International Publishing.

[39] K. Hamilton, S.-I. Shih, S. Mohammed, The development and validation of the rational and intuitive decision styles scale, Journal of Personality Assessment 98 (2016) 523–535.

[40] F. Costa, S. Ouyang, P. Dolog, A. Lawlor, Automatic generation of natural language explanations, in: Proceedings of the 23rd International Conference on Intelligent User Interfaces Companion, 2018, p. 57:1–57:2.

[41] D. C. Hernandez-Bocanegra, T. Donkers, J. Ziegler, Effects of argumentative explanation types on the perception of review-based recommendations, in: Adjunct Proceedings of the 28th ACM Conference on User Modeling, Adaptation and Personalization (UMAP ’20 Adjunct), 2020.

[42] P. J. Nelson, Consumer information and advertising, in: Economics of Information, 1981, p. 42–77.

[43] L. Klein, Evaluating the potential of interactivemedia through a new lens: Search versus experience goods, in: Journal of Business Research, volume 41, 1998, p. 195–203.