Discourse-based answering of *why*-questions

Employing RST structure for finding answers to *why*-questions

Suzan Verberne — Lou Boves — Peter-Arno Coppen — Nelleke Oostdijk

Department of Linguistics, Radboud University Nijmegen
s.verberne|l.boves|p.a.coppen|n.oostdijk@let.ru.nl

ABSTRACT. *This paper presents the work that we have carried out in investigating the purpose of discourse structure for* why*-question answering (*why*-QA). We developed a system for answering* why*-questions that employs the discourse relations in a pre-annotated document collection (the RST Treebank). With this method, we obtain a recall of 53.3% with a mean reciprocal rank (MRR) of 0.662. We argue that the maximum recall that can be obtained from the use of RST relations as proposed in the present paper is 58.0%. If we discard the questions that require world knowledge, maximum recall is 73.9%. We conclude that discourse structure can play an important role in complex question answering, but that more forms of linguistic processing are needed for increasing recall.*

RÉSUMÉ. *Cet article présente la recherche que nous avons réalisée en examinant le but de la structure du discours de réponse aux questions du type* pourquoi *(*why*-QA). Nous avons développé un système de réponse aux questions* pourquoi *qui utilise les relations RST dans une collection de documents pré-annotés (le RST Treebank). En appliquant cette méthode, nous obtenons un rappel de 53,3 % avec une Moyenne du Rang Inverse (MRR) de 0,662. Nous soutenons que le rappel maximum qui puisse être obtenu en utilisant les relations RST est de 58,0 %. En supprimant les questions qui requièrent une connaissance du monde, le rappel maximal serait de 73,9 %. Nous concluons que les structures du discours peuvent jouer un rôle important dans la réponse aux questions complexes, mais que l'augmentation du rappel nécessite davantage de sortes de traitements linguistiques.*

KEYWORDS: *Question answering,* why*-questions, discourse structure, RST*

MOTS-CLÉS : *Répondre questions, questions pourquoi, structure du discours, RST*

TAL Volume 47 – n°2/2007, pages 21 à 41

1. Introduction

Up to now, *why*-questions have largely been ignored by researchers in the field of question answering (QA). One reason for this is that the frequency of *why*-questions posed to QA systems is lower than that of other types of questions such as *who*- and *what*-questions (Hovy *et al.*, 2002). However, *why*-questions cannot be neglected: as input for a QA system, they comprise about 5 percent of all *wh*-questions (Hovy *et al.*, 2001; Jijkoun and De Rijke, 2005) and they do have relevance in QA applications (Maybury, 2003). A second reason why this type of question has largely been disregarded until now is that the techniques that have proven to be successful in QA for closed-class questions have been demonstrated to be not suitable for questions that expect an explanatory answer instead of a noun phrase (Kupiec, 1999).

Researchers in the field of discourse analysis have investigated whether knowledge about discourse structure can be put to use in a number of applications, among which language generation, text summarization, and machine translation (Carlson *et al.*, 2003). The relevance of discourse analysis for QA applications has been suggested by Marcu and Echihabi (2001) and Litkowski (2002). Breck *et al.* (2000) suggest that knowledge about discourse relations would have allowed their system for TREC-8 to answer *why*-questions. In this paper we take on the challenge and investigate to what extent discourse structure does indeed enable answering *why*-questions.

In the context of our research, a *why*-question is defined as an interrogative sentence in which the interrogative adverb *why* (or a synonymous word or phrase) occurs in (near) initial position. Furthermore, we only consider the subset of *why*-questions that could be posed to a QA system (as opposed to questions in a dialogue or in a list of frequently asked questions) and for which the answer is known to be present in some related document collection. In particular, our research is limited to questions obtained from a number of subjects who were asked to read documents from the collection and formulate *why*-questions that another person would be able to answer given the text.

The answer to a *why*-question is a clause or sentence (or a small number of coherent sentences) that answers the question without adding supplementary and redundant context. The answer is not necessarily literally present in the source document, but it must be possible to deduce it from the document.

An approach for automatically answering *why*-questions, like general approaches for factoid-QA, will involve at least four subtasks: (1) question analysis and query creation, (2) retrieval of candidate paragraphs or documents, (3) analysis and selection of text fragments, and (4) answer generation. In the current research, we want to investigate whether structural analysis and linguistic information can make QA for *why*-questions feasible. In previous work (Verberne, 2006), we focused on question analysis for *why*-questions. From other research reported on in the literature it appears that knowing the answer type helps a QA system in selecting potential answers. Therefore, we created a syntax-based method for the analysis of *why*-questions that was aimed at predicting the semantic answer type. We defined the following answer

types for *why*-questions, based on Quirk *et al.* (1985): *motivation*, *cause*, *circumstance* and *purpose*. Of these, *cause* (52%) and *motivation* (37%) are by far the most frequent types in our set of *why*-questions pertaining to newspaper texts. With our syntax-based method, we were able to predict the correct answer type for 77.5% of these questions (Verberne *et al.*, 2006b).

After analysis of the input question, the QA system will retrieve a small set of documents that possibly contain the answer. Analysis of the retrieved documents is then needed for extracting potential answers. Thus, a system for *why*-QA needs a text analysis module that yields a set of potential answers to a given *why*-question. Although we now have a proper answer type determination approach, the problem of answer extraction is still difficult. As opposed to factoid-QA, where named entity recognition can play an important role in the extraction of potential answers, finding potential answers to *why*-questions is still an unsolved problem. This means that we need to investigate how we can recognize the parts of a text that are potential answers to *why*-questions.

We decided to approach this answer extraction problem as a discourse analysis task. In this paper, we aim to find out to what extent discourse analysis can help in selecting answers to *why*-questions. We also investigate the possibilities of a method based on textual cues, and used that approach as baseline for evaluating our discourse-based method. Below, we will first introduce RST as a model for discourse analysis. Then we present our method for employing RST for *why*-QA, followed by the results that we obtained. We conclude this paper with a discussion of the limitations and possibilities of discourse analysis for the purpose of *why*-QA and the implications for future work.

2. Rhetorical Structure Theory (RST)

The main reasons for using RST as a model for discourse structure in the present research are the following. First, a treebank of manually annotated English texts with RST structures is available for training and testing purposes. This RST Discourse Treebank, created by (Carlson *et al.*, 2003), contains a selection of 385 Wall Street Journal articles from the Penn Treebank that have been annotated with discourse structure in the framework of RST. Carlson et al. adapted the default set of discourse relations proposed by Mann and Thompson for the annotation of the Wall Street Journal articles in the treebank. The annotations by Carlson et al. are largely syntax-based, which fits the linguistic perspective of the current research. A second reason for using RST is that relatively good levels of agreement have been measured between human annotators of RST, which indicates that RST analyses do not strongly depend on subjective interpretations of the structure of a text (Bosma, 2005).

In RST, the smallest units of discourse are called *elementary discourse units* (EDUs). In terms of the RST model, a rhetorical relation typically holds between two EDUs, one of which (the *nucleus*) is more essential for the writer's intention than

the other (the *satellite*). If two related EDUs are of equal importance, there is a *multinuclear relation* between them. Two or more related EDUs can be grouped together in a larger *span*, which in its turn can participate in another relation. By grouping and relating spans of text, a hierarchical structure of the text is created. In the remainder of this paper, we will refer to such a hierarchical structure as an *RST tree*.

3. Our method for discourse-based *why*-QA

3.1. *Main ideas and procedure*

Let us consider a *why*-question-answer pair and the RST structure of the corresponding source text. We hypothesize the following:

1. The question topic[1] corresponds to a span of text in the source document and the answer corresponds to another span of text;
2. In the RST structure of the source text, an RST relation holds between the text span representing the question topic and the text span representing the answer.

If both hypotheses are true, then RST can play an important role in answering *why*-questions.

For the purpose of testing our hypotheses, we need a number of RST annotated texts and a set of question-answer pairs that are linked to these texts. Therefore, we set up an elicitation experiment using the RST Treebank as data set. We selected seven texts from the RST Treebank of 350–550 words each. Then we asked native speakers to read one of these texts and to formulate *why*-questions for which the answer could be found in the text. The subjects were also asked to formulate answers to each of their questions. This resulted in a set of 372 *why*-question and answer pairs, connected to seven texts from the RST Treebank. On average, 53 question-answer pairs were formulated per source text. There is much overlap in the topics of the questions, as we will see later.

A risk of gathering questions following this method, is that the participants may feel forced to come up with a number of *why*-questions. This may lead to a set of questions that is not completely representative for a user's real information need. We believe however that our elicitation method is the only way in which we can collect questions connected to a specific (closed) set of documents. We will come back to the representativeness of our data collection in section 5.3.

We performed a manual analysis on 336 of the collected question-answer pairs in order to check our hypotheses – we left out the other (randomly selected) pairs for future testing purposes (not addressed in the current paper). We chose an approach

1. The topic of a *why*-question is the proposition that is questioned. A *why*-question has the form 'WHY P?', in which the proposition P is the topic. (Van Fraassen, 1988)

in which we analyzed our data according to a clear step-by-step procedure, which we expect to be suitable for answer extraction performed by a QA system. This means that our manual analysis will give us an indication of the upper bound of the performance that can be achieved using RST following the proposed approach.

First, we selected a number of relation types from Carlson *et al.*'s relation set, which we believed might be relevant for *why*-QA. We started with the four answer types mentioned in the introduction of this paper (cause, purpose, motivation and circumstance), but it soon appeared that there is no one-to-one relation between the four classes we defined based on Quirk *et al.* (1985) and relation types in Carlson *et al.*'s set. For instance, Carlson *et al.*'s relation set does not contain the relation type *motivation*, but uses *reason* instead. Moreover, we found that the set of relations to which at least one *why*-question in our data collection refers is broader than just *cause*, *circumstance*, *purpose* and *reason*. Therefore, we extended the list during the manual analysis. The final set of selected relations is shown in Table 1.

Table 1. *Selected relation types*

Cause	Circumstance	Condition
Elaboration	Explanation-argumentative	Evidence
Interpretation	List	Problem-Solution
Purpose	Reason	Result
Sequence		

For the majority of these relations, the span of text that needs explanation (or elaboration, evidence, etc.) is the nucleus of the relation, and the span of text giving this explanation is the satellite. The only exception to this rule is the cause relation, where the cause is given by the nucleus and its result by the satellite. Knowing this, we used the following procedure for analyzing the questions and answers:

I. Identify the topic of the question.

II. In the RST tree of the source document, identify the span(s) of text that express(es) the same proposition as the question topic.

III. Is the found span the nucleus of a relation of one of the types listed in Table 1 (or, in case of cause relations, the satellite)? If it is, go to IV. If it is not, go to V.

IV. Select the related satellite (or nucleus in case of a cause relation) of the found span as an answer.

V. Discard the current text span.

The effects of the procedure can best be demonstrated by means of an example. Consider the following question, formulated by one of the native speakers after he had read a text about the launch of a new TV channel by Whittle Communications L.P.

Q: Why does Christopher Whittle think that Channel One will have no difficulties in reaching its target?

The topic of this question is *Christopher Whittle thinks that Channel One will have no difficulties in reaching its target*. According to our first hypothesis, the proposition expressed by the question topic matches a span in the RST structure of the source document. We manually selected the following text fragment which expresses the proposition of the question topic:

> "What we've done in eight weeks shows we won't have enormous difficulties getting to the place we want to be", said Mr. Whittle.

This sentence covers span 18–22 in the corresponding RST tree, which is shown in Figure 1.

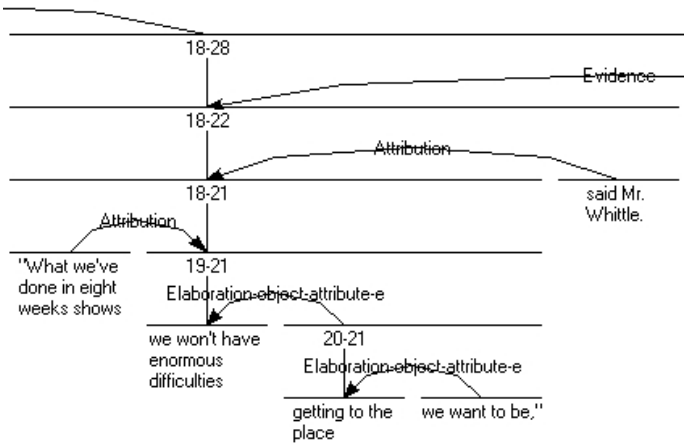

Figure 1. *RST sub-tree for the text span "What we've done in eight weeks shows we won't have enormous difficulties getting to the place we want to be, said Mr. Whittle."*

In this way, we tried to identify a span of text corresponding to the question topic for each of the 336 questions.

In cases where we succeeded in selecting a span of text in the RST tree corresponding to the question topic, we searched for potential answers following step III and IV from the analysis procedure. As we can see in Figure 1, the span *What we've done in eight weeks shows we won't have enormous difficulties getting to the place we want to be, said Mr. Whittle* is the nucleus of an evidence relation. Since we assumed that an evidence relation may lead to a potential answer (Table 1), we can select the satellite of this relation, span 23–28, as an answer (see Figure 2 below):

A: He said his sales force is signing up schools at the rate of 25 a day. In California and New York, state officials have opposed Channel One. Mr. Whittle said private and parochial schools in both states will be canvassed to see if they are interested in getting the programs.

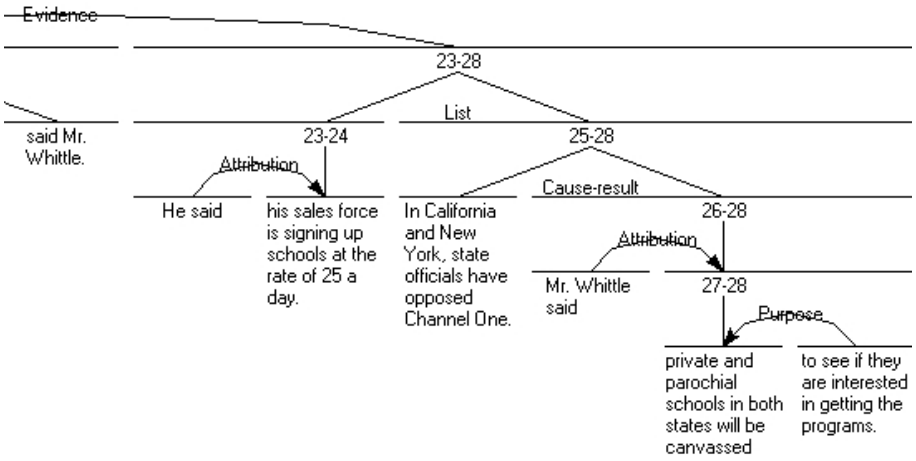

Figure 2. *RST sub-tree containing the satellite span "He said his sales force ... to see if they are interested in getting the programs."*

We analyzed all 336 *why*-questions following this procedure. The result of this manual analysis is a table containing all questions and for each question the following fields: (a) the manually identified topic from the source text with its corresponding span from the RST tree; (b) the answer span that we found for the question topic; (c) the type of relation that holds between topic span and answer span, if there is a relation; and (d) information about whether the answer found is correct. We will come back to this in section 4.1, where we discuss the outcome of the manual analysis.

3.2. *Implementation*

We implemented the procedure presented above in a Perl script. In section 3.1, we assumed that the RST structure can lead to a possible answer span once the topic span has been identified as a nucleus of a relevant relation. Therefore, the most critical task of our procedure is step II: to identify the span(s) of text that express(es) the same proposition as the question topic.

Since we are only interested in those spans of text that participate in an RST relation (step III), we need a list of all nuclei and satellites for each document in our data collection, so that our system can select the most relevant nuclei. Therefore, we built an indexing script that takes as input file the RST structure of a document,

and searches it for instances of relevant relations (Table 1). It then extracts for each relation its nucleus, satellite and relation type and saves it to an index file (in plain text). In case of a multinuclear relation, the script saves both nuclei to the index file. Moreover, cause relations are treated a bit differently from the other relation types. In cause relations, as explained before, the span of text describing the cause is marked as nucleus, not as satellite. Thus, the satellite of cause relations should be indexed for matching to the question topic instead of the nucleus. Therefore, nucleus and satellite are transposed when indexing cause relations. Below, where we use the term *nucleus* in describing the retrieval process, we mean the satellite for cause relations and the nucleus for all other relations.

Figures 3 and 4 below illustrate the conversion from an RST structure file to an index file. We created indexes for all documents in the RST Treebank.

```
( Nucleus (span 29 32) (rel2par span)
 ( Nucleus (leaf 29) (rel2par span) (text _!that interior regions of Asia
                                          would be among the first_!) )
 ( Satellite (span 30 32) (rel2par elaboration-object-attribute-e)
  ( Nucleus (leaf 30) (rel2par span) (text _!to heat up in a global warming_!) )
  ( Satellite (span 31 32) (rel2par consequence-n-e)
   ( Nucleus (leaf 31) (rel2par span) (text _!because they are far from
                                                        oceans,_!) )
   ( Satellite (leaf 32) (rel2par elaboration-additional) (text _!which
                                  moderate temperature changes.<P>_!) )
  )
 )
)
```

Figure 3. *Fragment of the original RST structure*

```
> consequence
   1. Nucleus (30): to heat up in a global warming
   2. Satellite (31 32): because they are far from oceans, which
                                  moderate temperature changes

> elaboration
   1. Nucleus (31): because they are far from oceans
   2. Satellite (32): which moderate temperature changes
```

Figure 4. *Fragment of the resulting index*

For the actual retrieval task, we wrote a second Perl script that takes as input one of the document indices, and a question related to the document. Then it performs the following steps:

1. Read the index file and normalize each nucleus in the index. Normalization includes at least removing all punctuation from the nucleus. Other forms of normalizing that we explored are lemmatization, applying a stop list, and adding synonyms for each content word in the nucleus. These normalization forms are combined into a number of configurations, which are discussed in section 4.2;

2. Read the question and normalize it, following the same normalization procedure as for the nuclei;

3. For each nucleus in the index, calculate the likelihood P (Nucleus | Question) using the following language model (N = nucleus; Q = question; R = relation type for nucleus):

   Nucleus likelihood $P(N|Q) \sim P(Q|N) \cdot P(N)$

   Question likelihood $P(Q|N) = \frac{\#\ question\ words\ in\ nucleus}{\#\ words\ in\ nucleus}$

   Nucleus Prior $P(N) = \frac{1}{\#\ nuclei\ in\ document} \cdot P(R)$

   Relation Prior $P(R) = \frac{\#\ instances\ of\ this\ relation\ type\ in\ question\ set}{\#\ occurrences\ of\ this\ relation\ type\ in\ treebank}$

   For calculation of the relation prior P(R), we counted the number of occurrences of each relation type in the complete RST Treebank. We also counted the number of occurrences to which at least one question in our data collection refers. The proportion between these numbers, the relation prior, is an indication of the relevance of the relation type for *why*-question and answer pairs.
   For convenience, we take the logarithm of the likelihood. This avoids underflow problems with very small probabilities. Thus, since the range of the likelihood is [0..1], the range of the logarithm of the likelihood is [-∞..0];

4. Save all nuclei with a likelihood greater than the predefined threshold (see section 4.2);

5. Rank the nuclei according to their likelihood;

6. For each of the nuclei saved, print the corresponding answer satellite and the calculated likelihood.

We measured the performance of our implementation by comparing its output to the output of the manual analysis described in section 3.1.

4. Results

In this section, we will first present the outcome of the manual analysis, which gives an indication of the performance that can be achieved by a discourse-based system for *why*-QA (section 4.1).

Then we present the performance of the current version of our system. When presenting the results of our system, we can distinguish two types of measurements. First, we can measure the system's absolute quality in terms of recall and mean reciprocal rank (MRR). Second, we can measure its performance relative to the results we obtained from the manual analysis. In this section, we do both (section 4.2).

4.1. *Results of the manual analysis*

As described in section 3.1, our manual analysis procedure consists of four steps: (I) identification of the question topic, (II) matching the question topic to a span of text, (III) checking whether this span is the nucleus of an RST relation (or satellite, in case of a cause relation), and (IV) selecting its satellite as answer. Below, we will discuss the outcome of each of these sub-tasks.

The first step succeeds for all questions, since each *why*-question has a topic. For the second step, we were able to identify a text span in the source document that represents the question topic for 279 of the 336 questions that we analyzed (83.0%). We found that not every question corresponds to a unique text span in the source document. For these 279 questions, we identified 84 different text spans. This means that on average, each text span that represents at least one question topic is referred to by 3.3 questions. For the other 57 questions, we were not able to identify a text span in the source document that represents the topic. These question topics are not explicitly mentioned in the text but inferred by the reader using world knowledge. We will come back to this in section 5.1.

For 207 of the 279 questions that have a topic in the text (61.6% of all questions), the question topic participates in a relation of one of the types in Table 1 (step III).

Evaluation of the fourth step, answer selection, needs some more explanation. For each question, we selected as an answer the satellite that is connected to the nucleus corresponding to the question topic. For the purpose of evaluating the answers found using this procedure, we compared them to the user-formulated answers. If the answer found matches at least one of the answers formulated by native speakers in meaning (not necessarily in form), then we judged the answer found as correct. For example, for the question *Why did researchers analyze the changes in concentration of two forms of oxygen?*, two native speakers gave as an answer *To compare temperatures over the last 10,000 years*, which is exactly the answer that we found following our procedure. Therefore, we judged our answer as correct, even though eight subjects gave a different answer to this question. Evaluating the answer that we found to the question *Why does Christopher Whittle think that Channel One will have no difficulties in reaching its target?* is slightly more difficult, since it is longer than any of the answers formulated by the native speakers. We got the following user-formulated answers for this question:

(1) Because schools are subscribing at the rate of 25 a day.

(2) Because agents are currently signing up 25 schools per day.

(3) He thinks he will succeed because of what he has been able to do so far.

(4) Because of the success of the previous 8 weeks.

Answers 1 and 2 refer to leaf 24 in the RST tree (see Figure 2); answers 3 and 4 refer to leaf 18 in the tree (see Figure 1). None of these answers correspond exactly to

the span that we found as answer using the answer extraction procedure (*He said his sales force ... in getting the programs.*). However, since some of the user-formulated answers are part of the answer span found, and because the answer is still relatively short, we judged the answer found as correct.

We found that for 195 questions, the satellite connected to the nucleus corresponding to the topic is a correct answer. This is 58.0% of all questions.

The above figures are summarized in Table 2.

Table 2. *Outcome of manual analysis*

Question	# questions	% of questions
Questions analyzed	336	100
Questions for which we identified a text span corresponding to the topic	279	83.0
Questions for which the topic corresponds to the nucleus of a relation (or satellite in case of a cause relation)	207	61.6
Questions for which the satellite of this relation is a correct answer	195	58.0

In section 5.1, we will come back to the set of questions (42%) for which our procedure did not succeed.

4.2. *System evaluation*

We evaluate our system using the outcome of our manual analysis as reference. We used the answer that we found during manual analysis as reference answer. We measured recall (the proportion of questions for which the system gives at least the reference answer) and MRR (1/rank of the reference answer, averaged over all questions.) We also measured recall as proportion of the percentage of questions for which the manual analysis led to the correct answer (58%, see Table 2 above).

We tested a number of configurations of our system, in which we varied the following variables:

1. Applying a stop list to the indexed nuclei, i.e. removing occurrences of 251 high-frequent words, mainly function words;
2. Applying lemmatization, i.e. replacing each word by its lemma if it is in the CELEX lemma lexicon (Baayen *et al.*, 1993). If it is not, the word itself is kept;
3. Expanding the indexed nuclei with synonym information from WordNet (Fellbaum, 1998), i.e. for each content word in the nucleus (nouns, verbs and adjectives), searching the word in WordNet and adding to the index all lemmas from its synonym set;

4. Changing weights between stop words and non-stop words.

We found that best performing is the configuration in which stop words are not removed, lemmatization is applied, no synonyms are added, and stop words and non-stop words are weighted 0.1/1.9. Moreover, in order to reduce the number of answers per question, we added a threshold to the probability of the nuclei found. For deciding on this threshold, we investigated what the log probability is that our system calculates for each of the correct (reference) answers in our data collection. As threshold, we chose a probability that is slightly lower than the probabilities of these reference answers.

For measuring the performance of our system, we added a function to the system in Perl that compares the answer spans found by the system to the answer in the reference table that was manually created (see section 3.1). We ran our system on the 336 questions from our data collection.

With the optimal configuration as described above, the system found the reference answer for 179 questions. So, the system obtains a recall of 53.3% (179/336). This is 91.8% of the questions for which the RST structure led to the correct answer in the manual analysis (179/195). The average number of answers that the system gives per question is 16.7. The mean reciprocal rank for the reference answer is fairly high: 0.662. For 29.5% of all questions, the reference answer is ranked in first position. This is 55.3% of the questions for which the system retrieved the reference answer.

An overview of the system results is given in Tables 3 and 4 below.

We should note here that recall will go up if we add synonyms to the index for all nuclei, but this lowers MRR and heavily slows down the question-nucleus matching process.

Table 3. *Main results for optimal configuration*

Recall (%)	53.3
Recall as proportion of questions for which the RST structure can lead to a correct answer (%)	91.8
Average number of answers per question	16.7
Mean reciprocal rank	0.662

Table 4. *Ranking of reference answer*

Answer rank	# questions	% of questions
Reference answer found	179	53.3
Reference answer ranked in 1st position	99	55.3
Reference answer ranked in 2nd to 10th position	60	33.5
Reference answer ranked in other position	20	11.2
Reference answer not found	157	46.7

5. Discussion of the results

In the discussion of the results that we obtained, we will focus on two groups of questions. First, we will discuss the questions for which we could not find an answer in our manual analysis following the procedure proposed (procedure shortcomings). Second, we will consider the questions for which we found an answer using manual analysis but our system could not find this answer (system shortcomings). For both groups of questions, we will study the cases for which we did not succeed, and make recommendations for future improvements of our system. In the last part of this section, we will give an overview of the types of RST relations that were found to play a role in *why*-QA.

5.1. *Discussion of procedure shortcomings*

5.1.1. *Error analysis*

We reported in section 4.1 that for 195 *why*-questions (58.0% of all questions), the answer could be found after manually matching the question topic to the nucleus of an RST relation and selecting its satellite as answer. This means that for 141 questions (42.0%), our method did not succeed. We distinguish four categories of questions for which we could not extract a correct answer using this method (percentages are given as part of the total of 336 questions):

1. Questions whose topics are not or only implicitly supported by the source text (57 questions, 17.0%). Half of these topics is supported by the text, but only implicitly. The propositions underlying these topics are true according to the text, but we cannot denote a place in the text where this is confirmed explicitly. Therefore, we were not able to select a span corresponding to the topic. For example, the question *Why is cyclosporine dangerous?* refers to a source text that reads *They are also encouraged by the relatively mild side effects of FK-506, compared with cyclosporine, which can cause renal failure, morbidity, nausea and other problems.* We can deduce from this text fragment that cyclosporine is dangerous, but we need knowledge of the world (*renal failure, morbidity, nausea and other problems are dangerous*) to do this. For the other half of these questions, the topic is not supported at all by the text, even not implicitly. For example, *Why is the initiative likely to be a success?*, whereas nowhere in the text there is evidence that the initiative is likely to be a success.

2. Questions for which both topic and answer are supported by the source text but there is no RST relation between the span representing the question topic span and the answer span (55 questions, 16.4%). In some cases, this is because the topic and the answer refer to the same EDU. For example, the question *Why were firefighters hindered?* refers to the span *Broken water lines and gas leaks hindered firefighters' efforts,* which contains both question topic and answer. In

other cases, question topic and answer are embedded in different, non-related spans, which are often remote from each other.

3. Questions for which the correct (i.e. user-formulated) answer is not or only implicitly supported by the text (17 questions, 5.1%). In these cases, the question topic is supported by the text, but we could not find evidence in the text that the answer is true or we are not able to identify the location in the text where it is confirmed explicitly. For example, the topic of the question *Why was Gerry Hogan interviewed?* corresponds to the text span *In an interview, Mr. Hogan said*. The native speaker that formulated this question gave as answer *Because he is closer to the activity of the relevant unit than the Chair, Ted Turner, since he has the operational role as President.* The source text does read that Mr. Hogan is president and that Ted Turner is chair, but the assumption that Gerry Hogan is closer to the activity than Ted Turner has been made by the reader, not by the text.

4. Questions for which the topic can be identified in the text and matched to the nucleus of a relevant RST relation, but the corresponding satellite is not suitable or incomplete as answer (12 questions, 3.6%). These are the questions that in table 2 make the difference between the last two rows (207-195). Some answers are unsuitable because they are too long. For instance, there are cases where the complete text is an elaboration of the sentence that corresponds to the question topic. In other cases, the answer satellite is incomplete compared to the user-formulated answers. For example, the topic of the question *Why did Harold Smith chain his Sagos to iron stakes?* corresponds to the nucleus of a circumstance relation that has the satellite *After three Sagos were stolen from his home in Garden Grove*. Although this satellite gives a possible answer to the question, it is incomplete according to the user-formulated answers, which all mention the goal *To protect his trees from thieves.*

Questions of category 1 above cannot be answered by a QA system that expects the topic of an input question to be present and identifiable in a closed document collection. If we are not able to identify the question topic in the text manually, then a retrieval system cannot either. A comparable problem holds for questions of category 3, where the topic is supported by the source text but the answer is not or only implicitly. If the system searches for an answer that cannot be identified in a text, the system will clearly not find it in that text. In the cases where the answer is implicitly supported by the source text, world knowledge is often needed for deducing the answer from the text, like in the examples of cyclosporine and Gerry Hogan above. Therefore, we consider the questions of types 1 and 3 as unsolvable by a QA system that searches for the question topic in a closed document collection. Together these categories cover 22.0% of all *why*-questions.

Questions of category 2 (16.4% of all questions) are the cases where both question topic and answer can be identified in the text, but where there is no RST relation between the span representing the question topic span and the answer span. We can

search for ways to extent our algorithm so that it can handle some of the cases mentioned. For instance, we can add functionality for managing question-answer relations on sub-EDU level. We think that in some of these cases, syntactic analysis can help in extracting the relation from the EDU. The example question above, *Why were firefighters hindered?* can be answered by a QA system if it knows that the question can be rephrased by *What hindered firefighters?*, and that has syntactic information about the EDU *Broken water lines and gas leaks hindered firefighters' efforts*. The risk of adding functionality for cases like this is that the number of possible answers per question will increase, decreasing the MRR. We should investigate to what extent syntactic analysis can help in cases where the answer lies in the same EDU as the question. For cases where question topic and answer are embedded in non-related spans, we can at the moment not propose smart solutions that will increase recall without heavily decreasing the MRR. The same holds for questions of category 4 (3.3%), where RST leads to an answer that is incomplete or unsuitable.

We can conclude from this analysis that there is a subset of *why*-questions (22.0%) that cannot be answered by a QA system that uses a closed document collection since knowledge of the world is essential for answering these questions. Moreover, there is a further subset of *why*-questions (16.4% + 3.6%) that cannot be answered by a system that uses RST structure only, following the approach that we proposed. Together, this means that 42.0% of *why*-questions cannot be answered following the suggested approach. Thus, the maximum recall that can be achieved with this method is 58.0%. If we discard the 72 (57+15) questions that require world knowledge, maximum recall would be 73.9% (195/(336-72)).

5.1.2. *Comparison to baseline*

In order to judge the merits of RST structure for why-QA, we investigated the possibilities of a method based on textual cues (without discourse structure). To that goal, we analyzed the text fragments related to each question-answer pair in our data collection. For each of these pairs, we identified the item in the text that indicates the answer. For 50% of the questions, we could identify a word or group of words that in the given context is a cue for the answer. Most of these cues, however, are very frequent words that also occur in many non-cue contexts. For example, the subordinator *that* occurs 33 times in our document collection, only 3 of which are referred to by one or more *why*-questions. This means that only in 9% of the cases, the subordinator *that* is a *why*-cue. The only two words for which more than 50% of the occurrences are *why*-cues, are *because* (for 4.5% of questions) and *since* (2.2%). Both are a *why*-cue in 100% of their occurrences. For almost half of the question-answer pairs that do not have an explicit cue in the source text, the answer is represented by the sentence that follows (17.6% of questions) or precedes (2.8%) the sentence that represents the question.

Having this knowledge on the frequency of cues for *why*-questions, we defined the following baseline approach:

I. Identify the topic of the question.

II. In the source document, identify the clause(s) that express(es) the same proposition as the question topic.

III. Does the clause following the matched clause start with *because* or *since*? If it does, go to IV. If it does not, go to V.

IV. Select the clause following the matched clause as answer.

V. Select the sentence following the sentence containing the matched clause as answer.

A system that follows this baseline method can obtain a maximum recall of 24.3% (4.5+2.2+17.6). This means that an RST-based method can improve recall by almost 140% compared to a simple cue-based method (58.0% compared to 24.3%).

5.2. *Discussion of system shortcomings*

There are 22 questions for which the manual analysis led to a correct answer, but the system did not retrieve this reference answer. For 17 of them, the nucleus that was matched to the question topic manually, is not retrieved by the system because there is no (or too little, given the threshold) lexical overlap between the question and the nucleus that represents its topic. For example, the question *Why are people stealing cycads?* can be matched manually to the span *palm-tree rustling is sprouting up all over Southern California*, but there are no overlapping words. If we add synonyms to our index for each nucleus (see section 3.2), then 10 of these questions can be answered by the system, increasing recall.

For three other questions, it is our algorithm that fails: these are cases where the question topic corresponds to the satellite of an elaboration relation, and the answer to the nucleus, instead of vice versa. We implemented this functionality for cause relations (see section 4.2), but implementing it for elaboration relations, where these topic-satellite correspondences are very rare, would increase the number of answers per questions and decrease MRR without increasing recall very much.

5.3. *RST relations that play a role in* **why**-*QA*

We counted the number of occurrences of the relation types from Table 1 for the 195 questions where the RST relation led to a correct answer. This distribution is presented in Table 5. The meaning of the column *Relative frequency* in this context will be explained below.

As shown in table 5, the relation type with most referring question-answer pairs, is the very general elaboration relation. It seems striking that *elaboration* is more

Table 5. *Addressed relation types*

Relation type	# referring questions	Relative frequency
Means	4	1.000
Purpose	28	0.857
Consequence	30	1.000
Evidence	7	0.750
Reason	19	0.750
Result	19	1.000
Explanation-argumentative	14	0.571
Cause	7	0.500
Condition	1	0.333
Interpretation	7	0.333
Circumstance	1	0.143
Elaboration	53	0.112
Sequence	1	0.091
List	4	0.016
Problem-Solution	0	0.000

frequent as a relation between a *why*-question and its answer than *reason* or *cause*. However, if we look at the relative frequency of the addressed relation types, we see another pattern: in our collection of seven source texts, *elaboration* is a very frequent relation type. In the seven texts that we consider, there are 143 occurrences of an elaboration relation. Of the 143 nuclei of these occurrences, 16 were addressed by one or more *why*-questions, which gives a relative frequency of around 0.1. *Purpose*, on the other hand, has only seven occurrences in our data collection, six of which being addressed by one or more questions, which gives a relative frequency of 0.857. *Reason* and *evidence* both have only four occurrences in the collection, three of which have been addressed by one or more questions. *Consequence* even has a relative frequency of 1.000

The table shows that if we address the problem of answer selection for *why*-questions as a discourse analysis task, the range of relation types that can lead to an answer is broad and should not be implemented too rigidly.

In section 3.1, we pointed out that our data collection may not be fully representative of a user's information need, due to our elicitation method using a closed document set. The relation types in table 5 confirm that assumption to some extent: the presence of relation types such as *means* and *condition* suggests that the subjects in some cases formulated *why*-questions whereas they would have formulated *how*- or *when*-questions in case of an actual information need. A question-answer pair like *Why could FK-506 revolutionize the organ transplantation field? - Because it reduces harmful side effects and rejection rates*, whereas the text reads *FK-506 could revolutionize the transplantation field by reducing harmful side effects* exemplifies this.

If we want to know our system's performance on *why*-questions that are representative for a user's information need, we are interested in those questions whose answers can be found through a 'core-*why* relation' like *cause* and *reason*.

If we only consider the relation types that have relative frequency higher than or equal to 0.5, we see that these relation types are in general closer to the concept of *reason* as general answer type of *why*-questions (Verberne, 2006) than the relation types with a relative frequency lower than 0.5. We also see that the most frequent answer types that we defined for question analysis (see section 1) come back in this set of relation types. *Purpose* and *reason*, as defined by Carlson and Marcu (2001), correspond to our definition of the answer type *motivation* (Verberne *et al.*, 2006a). Carlson and Marcu (2001)'s *consequence*, *result* and *cause* relations can, based on their definitions, be grouped together as our answer type *cause*.

We investigated to what extent the performance of our system depends on the type of relation that leads from question topic to the reference answer. For this purpose, we split the relation types found in two categories:

- Relation types that are conceptually close of the general answer type *reason* ('core-*why* relations'): *Purpose*, *Consequence*, *Evidence*, *Reason*, *Result*, *Explanation-argumentative* and *Cause*. These relation types all have a relative frequency higher than 0.5 for *why*-questions.
- Relation types that are less applicable to *why*-questions ('non-*why* relations'): *Means*, *Condition*, *Interpretation*, *Circumstance*, *Elaboration*, *Sequence*, *List* and *Problem-Solution*.

We considered the set of 207 questions for which the topic corresponds to the nucleus of a relation (thereby excluding the 74 questions whose topic or answer is unsupported, or where the RST relation does not lead to an answer) and measured our system's recall on this set of questions. This is 77.5% — which is higher than the total recall of 51.2% because we excluded the majority of problematic cases. We then split the set of 207 questions into one set of questions whose answers can be found through a core-*why* relation (130 questions), and one set of questions that correspond to a non-*why* relation (77 questions) and ran our system on both these sets. For the core-*why* relation types, we found a system recall of 88.5% and for the non-*why* relation types a system recall of 60.3%. Moreover, we found that the remaining 11.5% for the core-*why* relation types suffer from lexical matching problems (see section 5.2) instead of procedural problems: for 100% of these questions, the satellite of the relation is a correct answer. For the non-*why* relation types, this is 85.9%.

Another problem of our data collection method, is that the questions formulated by the readers of the text (in particular the questions relating to core-*why* relations) will probably be influenced by the same linguistic cues that are used by the annotators that built the RST structures: cue phrases (like *because* denoting an explanation relation) and syntactic constructions (like infinite clauses denoting a purpose relation). This is an unwelcome correlation, since in a working QA system users will not have access

to the documents. Future work should indicate to what extent questions representing a real information need refer to *why*-relations in the RST structure.

6. Conclusions

We created a method for *why*-QA that is based on discourse structure and relations. The main idea of our approach is that the propositions of a question topic and its answer are both represented by a text span in the source text, and that an RST relation holds between these spans. A *why*-question can then be answered by matching its topic to a span in the RST tree and selecting the related span as answer.

We first investigated the possible contribution of the current RST approach to *why*-QA by performing a manual analysis of our set of 336 questions and answers collected through elicitation from native speakers and connected to seven RST-annotated texts. From the evaluation of our manual analysis, we concluded that for 58.0% of our *why*-questions, an RST relation holds between the text span corresponding to the question topic and the text span corresponding to the answer.

We implemented this method for discourse-based *why*-QA using the RST Treebank as document collection. Our system obtains a recall of 53.3% (91.8% of the manual score) with a MRR of 0.662.

In section 5.1, we conclude from the analysis of procedure shortcomings that there is a subset of *why*-questions (22.0%) that cannot be answered by a QA system that expects the topic of an input question to be present and identifiable in a closed document collection. For these questions, either the topic or the user-formulated answer is not or only implicitly supported by the corresponding source text, which means that world knowledge is necessary for answering these questions. Furthermore, there is a further subset of *why*-questions (16.4%) that cannot be answered by a system that uses RST structure following the approach we proposed. For these questions, there is no RST relation between the span corresponding to the question topic and the span corresponding to its answer. A third subset (3.6%) of problematic questions contains those questions for which RST leads to an unsuitable or incomplete answer. Together, this means that 42.0% of *why*-questions cannot be answered following the suggested approach. Thus, the maximum recall that can be achieved with this method is 58.0%. If we discard the questions that require world knowledge, maximum recall would be 73.9%. An even higher performance can be achieved if we would only consider those questions that refer to core-*why* relations in the text like *cause* and *reason*.

In the near future we will focus our research on three topics. Firstly, we will investigate the *why*-questions (16.4% of the questions in our collection) where both topic and answer are supported by the source text, but where there is no RST relation between the span representing the question topic span and the answer span. We think that other types of linguistic analysis, or different exploitation of the RST structure can help for answering these questions.

Secondly, we aim to create and annotate a test corpus connected to *why*-questions that originate from real users' information needs, based on the *why*-questions collected for the Webclopedia project (Hovy *et al.*, 2002). With this set, we will investigate first to what extent questions representing real information needs refer to *why*-relations in a document's RST structure and second what the performance of our method is on such a set of questions.

Thirdly, we should note that in a future application of *why*-QA using RST, the system will not have access to a manually annotated corpus—it has to deal with automatically annotated data. We assume that automatic RST annotations will be less complete and less precise than the manual annotations are. As a result of that, performance would decline if we were to use automatically created annotations. Some work has been done on automatically annotating text with discourse structure. Promising is the done work by Marcu and Echihabi (2001), Soricut and Marcu (2003) and Huong and Abeysinghe (2003). We plan to investigate to what extent we can achieve partial automatic discourse annotations that are specifically equipped to finding answers to *why*-questions. We think we can make such annotations feasible if we focus on the information that is needed for answering *why*-questions, based on the knowledge that we obtained from the work described in the present paper.

7. References

Baayen R., Piepenbrock R., van Rijn H., "The CELEX Lexical Database (CD-ROM)", *Linguistic Data Consortium, University of Pennsylvania, Philadelphia, PA*, 1993.

Bosma W., "Query-Based Summarization Using Rhetorical Structure Theory", *in* T. van der Wouden, M. Poß, H. Reckman, C. Cremers (eds), *15th Meeting of CLIN*, LOT, Leiden, p. 29-44, December, 2005. ISBN=90-76864-91-8.

Breck E., Burger J., Ferro L., House D., Light M., Mani I., "A Sys Called Qanda", *Proceedings of the Eighth Text REtrieval Conference (TREC 8)*, 2000.

Carlson L., Marcu D., *Discourse Tagging Reference Manual*, Univ. of Southern California/Information Sciences Institute. 2001.

Carlson L., Marcu D., Okurowski M. E., "Building a Discourse-Tagged Corpus in the Framework of Rhetorical Structure Theory", *in* J. van Kuppevelt, R. Smith (eds), *Current Directions in Discourse and Dialogue*, Kluwer Academic Publishers, p. 85-112, 2003.

Fellbaum C. E. (ed.), *WordNet: An Electronic Lexical Database*, Cambridge, Mass., MIT Press, 1998.

Hovy E., Gerber L., Hermjakob U., Lin C.-J., Ravichandran D., "Toward Semantics-Based Answer Pinpointing", *Proceedings of the DARPA Human Language Technology Conference (HLT)*, San Diego, CA, 2001.

Hovy E., Hermjakob U., Ravichandran D., "A Question/Answer Typology with Surface Text Patterns", *Proceedings of the Human Language Technology conference (HLT)*, San Diego, CA, 2002.

Huong T. L., Abeysinghe G., "A Study to Improve the Efficiency of a Discourse Parsing System", *Proceedings of CICLing-03*, Springer, p. 104-117, 2003.

Jijkoun V., De Rijke M., "Retrieving Answers from Frequently Asked Questions Pages on the Web", *Proceedings of CIKM-2005*, 2005.

Kupiec J., "MURAX: Finding and Organizing Answers from Text Search", *in* T. Strzalkowski (ed.), *Natural Language Information Retrieval*, Kluwer Academic Publishers, Dordrecht, Netherlands, p. 311-332, 1999.

Litkowski K., "CL Research Experiments in TREC-10 Question Answering", *The Tenth Text Retrieval Conference (TREC 2001). NIST Special Publication*, p. 500-250, 2002.

Marcu D., Echihabi A., "An Unsupervised Approach to Recognizing Discourse Relations", *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, Association for Computational Linguistics, Morristown, NJ, USA, p. 368-375, 2001.

Maybury M. T., "Toward a Question Answering Roadmap.", *in* M. T. Maybury (ed.), *New Directions in Question Answering*, AAAI Press, p. 8-11, 2003.

Quirk R., Greenbaum S., Leech G., Svartvik J., *A comprehensive grammar of the English language*, London, Longman, 1985.

Soricut R., Marcu D., "Sentence level discourse parsing using syntactic and lexical information", *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology—Volume 1*, Association for Computational Linguistics, Morristown, NJ, USA, p. 149-156, 2003.

Van Fraassen B., "The Pragmatic Theory of Explanation", *in* J. Pitt (ed.), *Theories of Explanation*, Oxford University Press, p. 135-155, 1988.

Verberne S., "Developing an Approach for *Why*-Question Answering", *Conference Companion of the 11th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2006)*, Trento, Italy, p. 39-46, 2006.

Verberne S., Boves L., Oostdijk N., Coppen P., "Data for question answering: the case of *why*", *Proceedings of the 5th edition of the International Conference on Language Resources and Evaluation (LREC 2006)*, Genoa, Italy, 2006a.

Verberne S., Boves L., Oostdijk N., Coppen P., "Exploring the use of linguistic analysis for *why*-question answering", *Proceedings of CLIN 2005*, Amsterdam, 2006b.