Data for question answering: The case of *why*

Suzan Verberne, Lou Boves, Nelleke Oostdijk, and Peter-Arno Coppen

Department of Linguistics, Radboud University Nijmegen
P.O. Box 9103, 6500 HD Nijmegen, The Netherlands
E-mail: s.verberne|l.boves|p.coppen|n.oostdijk@let.ru.nl

Abstract

For research and development of an approach for automatically answering *why*-questions (*why*-QA) a data collection was created. The data set was obtained by way of elicitation and comprises a total of 395 *why*-questions. For each question, the data set includes the source document and one or two user-formulated answers. In addition, for a subset of the questions, user-formulated paraphrases are available. All question-answer pairs have been annotated with information on topic and semantic answer type. The resulting data set is of importance not only for our research, but we expect it to contribute to and stimulate other research in the field of *why*-QA.

1. Introduction

Until now, research in the field of automatic question answering (QA) has focused on factoid (closed-class) questions like *who*-, *what*-, *where*- and *when*-questions. Results reported for the QA track of the Text Retrieval Conference (TREC) show that these types of *wh*-questions can be handled rather successfully (Voorhees 2003). In the current project, we aim at developing an approach for automatically answering *why*-questions (*why*-QA). So far, *why*-questions have largely been ignored by researchers in the QA field. One reason for this is that the frequency of *why*-questions in a QA context is lower than that of other questions like *who*- and *what*-questions (Hovy et al., 2002a). However, although *why*-questions are less frequent than some types of factoids (*who*, *what* and *where*), their frequency is not negligible: in a QA context, they comprise about 5 percent of all *wh*-questions (Hovy, 2001; Jijkoun, 2005) and they do have relevance in QA applications (Maybury, 2003).

In the current research, we want to investigate whether structural linguistic information and analysis can make QA for *why*-questions feasible. An approach for automatically answering *why*-questions will involve at least four subtasks: (1) question analysis and query creation, (2) retrieval of candidate paragraphs or documents, (3) paragraph analysis and selection, and (4) answer generation. In our research we will focus on the possible contributions of rule-based parsing to question and paragraph analysis.

In research in the field of QA, data sources of questions and answers play an important role. Appropriate data collections are necessary for the development and evaluation of QA systems (Voorhees, 2000). While in the context of the QA track of TREC data collections for factoid questions have been created, so far, no resources have been created for *why*-QA. For the purpose of the present research therefore, we have developed a data collection comprising a set of questions and corresponding answers.

2. Data for *why*-QA

In this section, we first describe the requirements that a data set must meet in order to be appropriate for research and development of an approach for *why*-QA (Section 2.1). We then discuss a number of existing sources of *why*-questions and we conclude that no existing set can fulfill the needs of our research (Section 2.2). Therefore, we decided to create a new data set of *why*-questions and corresponding answers which is specifically geared to the needs of *why*-QA. We describe the method that we employed in collecting the data set (Section 2.3).

2.1. Requirements for the data collection

The first requirement for an appropriate data set concerns the nature of the questions. In the context of the current research, a *why*-question is defined as an interrogative sentence in which the interrogative adverb *why* (or one of its synonyms) occurs in (near) initial position. Furthermore, we only consider the subset of *why*-questions that could be posed in a QA context and for which the answer is known to be present in some related document set. This means that our data set should only comprise *why*-questions for which the answer can be found in a fixed collection of documents. Also, the topic of the question itself should be present in one of the source texts. 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. This proposition should be true according to the document collection; otherwise, the question '*Why* P' cannot be answered.

Secondly, the data set should not only contain questions, but also the corresponding answers and source documents. The answer to a *why*-question is a clause or sentence (or a small number of coherent sentences) that answers the question without giving supplementary context. The answer is not necessarily literally present in the source document, but it must be possible to deduce from the document without the need for involving world knowledge not expressed in the text. For example, a possible answer to the question Q, based on the source snippet S, is the answer A below:

```
Q: Why are 4300 additional teachers
   required?
S: The school population is due to rise by
   74,000, which would require recruitment
   of an additional 4,300 teachers, [...]
A: Because the school population is due to
   rise by 74,000.
```

Finally, the size of the data set should be large and rich enough so that it is reasonable to expect that it covers the

variation that occurs in *why*-questions in a QA context. We will come back to this in section 3.2.

2.2. Existing resources of *why*-questions

As stated in section 1, data sources of questions and answers play an important role in QA-research. Having specified the requirements for a data collection needed for research into *why*-QA, we consider a number of existing sources of *why*-questions that we may use in our research.

First of all, for *why*-questions from corpora like the British National Corpus (BNC, 2002), in which questions typically occur in spoken dialogues, the answers are not structurally available with the questions, nor are they extractable from a document that has been linked to the question. Therefore, corpus data like these are not suitable for research into *why*-QA. The same holds for the data collected for the Webclopedia project (Hovy et al., 2002a), in which neither the answers nor the source documents were included. The questions, however, were collected in an actual QA environment, and therefore they are representative for the set of questions that we consider in the current research.

One could also consider questions and answers from frequently asked questions (FAQ) pages, like the large data set collected by Valentin Jijkoun (Jijkoun, 2005). However, in FAQ lists, each question is followed by a piece of text that not only contains the answer, but also a substantial amount of additional information, which makes it difficult to determine what the correct answer should be.

The questions in the test collections from the TREC-QA track do contain links to the possible answers and the corresponding source documents. However, these collections contain too few *why*-questions (less than ten per edition) to qualify as a data set that is appropriate for developing *why*-QA.

Although the BNC, Jijkoun and Webclopedia sources do not meet our requirements regarding the availability of answers, the questions from these sources can be used for investigating the syntactic properties of *why*-questions. We will come back to this in section 3.2.

2.3. Procedure for data collection

Given the lack of suitable data, a data set geared to the needs of QA research into *why*-questions had to be compiled. In order to meet the requirements that were formulated in section 2.1, it would be best to collect questions posed in an operational QA environment. In fact, this is what the compilers of the TREC-QA test collections did: they extracted factoid and definition questions from search logs donated by Microsoft and AOL (TREC, 2003). Since we do not have access to comparable sources, we decided to revert to the procedure used in earlier TRECs, and imitate a QA environment in an elicitation experiment. We extended the conventional procedure by collecting user-formulated answers in order to investigate the range of possible answers to each question. We also added paraphrases of collected questions in order to extend the syntactic and lexical variation in the data collection.

In the elicitation experiment, ten native speakers of English were asked to read five texts from Reuters' *Textline Global News* (1989) and five texts from *The Guardian on CD-ROM* (1992). The texts were around 500 words each. The experiment was conducted over the Internet, using a web form and some CGI scripts. In order to have good control over the experiment, we registered all participants and gave them a code for logging in on the web site. Every time a participant logged in, the first upcoming text that he or she did not yet finish was presented. The participant was asked to formulate one to six *why*-questions for this text, and to formulate an answer to each of these questions. The participants were explicitly told that it was essential that the answers to their questions could be found in the text. After submitting the form, the participant was presented the questions already formulated by one of the other participants and he or she was asked to formulate an answer to these questions too. The collected data were saved in text format, grouped per participant and per source document, so that the source information is available for each question. The answers have been linked to the questions.

In this experiment, 395 questions and 769 corresponding answers were collected.

Although the set of 395 questions already contained some variation, we still felt that the phrasing of the questions might have been influenced by the wording of the source texts. In order to expand the lexical and syntactic variation in the set of questions, a second elicitation experiment was set up, in which five participants from the first experiment were asked to paraphrase some of the original *why*-questions. From the original data set, 166 unique questions were randomly selected. The participants formulated 211 paraphrases in total for these questions. This means that some questions have more than one paraphrase. The paraphrases were saved in a text file that includes the corresponding original questions with the pointers to the answers and the corresponding source documents.

3. The resulting data collection

We expect that the collected data set is large enough for research and development in *why*-QA. We assume that it is syntactically and semantically representative for the range of *why*-questions that are formulated in the context of a QA system.

As stated above, 395 questions and 769 corresponding answers were collected in the first elicitation experiment. The number of answers would have been twice the number of questions if all participants would have been able to answer all questions that were posed by another participant. However, for 21 questions (5.3%), the second participant was not able to answer the first participant's question. In some cases (2.3% of the total set), this was due to the fact that the proposition of the topic was untrue. For example, one of the participants in our elicitation experiment addressing a text on a conflict between Mr. Bocuse and McDonalds posed the following question:

```
Why is Mr. Bocuse seeking a settlement?
```

This question presupposes the truth of the topic *Mr. Bocuse is seeking a settlement*, which is not true according to the text, in which McDonalds seeks a settlement. Therefore, this question cannot be answered.

3.1. Question topics

As explained in section 2.1, the topic of the question is the proposition that is questioned – the event that needs explanation according to the questioner. Identifying the question's topic and matching it to an item (event, state, or action) in the text is a prerequisite for finding the answer. However, it is only a small part of the complete question answering process.

We grouped the collected questions according to their topic. Two questions have the same topic if they refer to the same in the text. For example, the following four questions were formulated in the elicitation experiment by four different participants:

Q: Why are classes likely to be even bigger in the autumn term?
Q: Why are classes likely to be even bigger in the upcoming autumn term?
Q: Why do the school councils believe that class sizes will grow even more this year?
Q: Why will classes get even bigger?

From these questions, we identified the topic shared by the four questions:

Classes are likely to be even bigger in the autumn term.

This topic refers to the following sentence in the source text:

The council said that indications about school admissions this September show that classes are likely to be even bigger in the autumn term.

By analysis of the 395 questions in our dataset, we identified 203 different topics, which gives an average of almost two questions per topic. In reality, however, a small number of topics were very frequently questioned, whereas many other topics were addressed only once. For fifteen topics, five or more questions were formulated. 127 topics only occur once in our data set, meaning that they were questioned by only one participant.

3.2. Lexical and syntactic variation

3.2.1. The syntactic form of *why*-questions

As stated in section 2.1, we define a *why*-question as an interrogative sentence in which the interrogative adverb *why* (or one of its synonyms) occurs in (near) initial position. For the automatic analysis of *why*-questions it is important to know the range of syntactic structures that can occur in *why*-questions.

We studied the *why*-questions in our own data collection and the questions from the resources described in section 2.2, i.e. the BNC, the FAQ-data collected by Valentin Jijkoun, and the questions from the Webclopedia project. Based on these data and the description of *wh*-questions by Quirk (1985: 11.14), we defined the default word order for *why*-questions as:

[COORDINATOR] [ADVERBIAL] WHY [ADVERBIAL] OPERATOR SUBJECT PREDICATION

in which the bracketed constituents are optional. The *why*-element is incidentally realized by the clause *why is it that*. The ADVERBIAL position can be realized by a modifying adverb like *then*, *so*, *well*, or *now*, or a subordinate clause. OPERATOR is an auxiliary or a form of *be* or *have*, optionally followed by a negator. SUBJECT and PREDICATION comprise the same constituents as in declarative sentences.

Starting from the default word order, three types of grammatical highlighting can be applied to *why*-questions: clefting (example 1 below), extraposition of the clause subject (2), topicalization (3), and existential *there* constructions (4)

(1) Why is it me who has to make all the changes?
(2) Why is it easy to buy a gun in Eastern Europe?
(3) Garmont, why does that name sound familiar?
(4) Why is there a debate about class sizes?

In a dialogue context, questions with an incomplete syntactic structure are frequent. The BNC gives many examples of *why*-questions that miss the subject (example 1 below), the verb phrase (2) or both (3):

(1) Why bother?
(2) Why Elizabeth Taylor?
(3) Why not?

Although frequent in dialogues, no questions with an incomplete syntactic structure occur in our set of 395 questions that we collected through elicitation. Grammatical highlighting is also rare: there are no instances of clefting and topicalization in our set of questions. Fourteen questions (3.5%) have an extraposed clause subject, and eleven questions (2.8%) contain an existential *there* construction. This means that 94.7% of the questions in our data collection have the default word order. In the Webclopedia set, which contains 805 *why*-questions asked to a running QA system, we found comparable figures: 94.0% of the questions have the default word order. Four questions (0.5%) miss the verb phrase (e.g. *Why all the talk about Fidel Castro recently?*); none miss the subject. There are no occurrences of clefting and topicalization in this set either. Six questions (0.7%) have an extraposed clause subject, and 38 questions (4.7%) contain an existential *there* construction.

Considering the syntactic variation that we found in both our own set of questions and the Webclopedia questions, we assume that the large majority of *why*-questions formulated in a QA-context have the default word order. Therefore, syntactic analysis of *why*-questions can be relatively straightforward.

3.2.2. Differences between source text and questions

As concluded above, the large majority of *why*-questions have the default syntactic structure at the top level: to a large extent, the order of the constituents is fixed. In order to investigate the difficulty of the task of matching a question to a source text, we studied the questions in more detail and compared them to the corresponding source texts.

We compared the lexical and syntactic form of each question to the lexical and syntactic form of the piece of text that represents the topic of the question. It appeared that the participants in our experiment did not often re-use

parts of the source text in the formulation of their questions. In fact, in almost all cases where there is an one-to-one relation between the topic of the question and a specific sentence in the source text, participants rephrased the topic. This means that despite the fact that in our experiment the participants had access to the source text while formulating the questions, the differences between question and source text are relatively large, just as they are in an actual QA situation, in which the questioner does not have access to the source text while formulating a question.

In order to know the types of analysis needed for matching a question to a source text, we investigated the types of rephrasing that occur between question and text. We found the following types of rephrasing:

Lexical rephrasing, for example

S: Class sizes in schools in England and Wales have risen *for the second year running.*

Q: Why *have* class sizes in England risen *again*?

Verb tense rephrasing, for example,

S: Class sizes in schools in England and Wales *have risen.*

Q: Why *have* class sizes in England and Wales *been rising*?

Omission of optional constituents, for example,

S: Class sizes *in schools in England and Wales* have risen.

Q: Why have class sizes risen?

Sentence structure rephrasing, e.g.

S: which would *require recruitment* of an additional 4,300 teachers.

Q: Why do more than 4,000 teachers *need to be recruited*?

We also considered the set of paraphrases, in which the same types of rephrasing occur. As expected, the differences between the paraphrases and the source sentences are slightly bigger than the differences between the original questions and the source sentences.

For each text, we extracted a type list (a list of unique tokens) from which we removed the function words. The reason for deriving a type list rather than a lemma list is that morphological form variants like verb tenses are also relevant for the variation between source text and question. We measured the lexical overlap between the questions and the source texts as the number of words that appear in both the question and the type list of the source text. The average relative lexical overlap (the number of overlapping words divided by the total number of words in the question) between original questions and source texts is 0.35; the average relative lexical overlap between paraphrases and source texts is 0.31. This small difference in lexical overlap supports our assumption that our question set contains enough variation to be representative for questions in a QA context (since an extra rephrasing step increases the lexical distance only slightly).

The list of rephrasing types above shows that a thesaurus or a semantic net like WordNet is a necessary requirement for solving the lexical differences. Morphological analysis is needed for matching the tenses of the same verb to each other. A precise match of question and source sentence still remains very difficult. However, in many cases it is possible to match a piece of text to a question, based on lexical similarity. In the example below, the question has very little in common with the source snippet, but the lemmas *canal* and *turn over* will probably still lead to this text in the document retrieval step.

S: Last week a former Reagan administration official caused a flurry of concern in Washington when he declared that the US will never turn the Panama Canal over to the Panamanian government if it is controlled by General Noriega.

Q: Why are conservatives in the US reluctant to start turning over control of the canal?

However, finding the relevant document is only a small step to the actual identification of the answer. Much more knowledge and analysis is needed for finding that a possible answer to this question is *Because Noriega is in power in Panama.*

3.3. The range of possible answers

For the development of an approach for *why*-QA, it is important to know the range of possible answers to *why*-questions. In order to get an idea of the possible answers we compared the answers that different participants formulated to the same question to each other. If the answers refer to the same item (event, state or action) in the text, then we judged them as equivalent. In some of these cases, the lexical and syntactic forms of the two answers were completely different but their meanings are equivalent. For example,

Q: Why do more than 4,000 teachers need to be recruited?

A1: In order to maintain the size of classes at their present level.

A2: To prevent the classes from becoming too big.

For 60% of the questions, both participants gave equivalent answers. For the other 40% of questions, the answers refer to different items from the source text. E.g.

Q: Why did Mr. Bush phone Mr. Tillotson when he was in Guatemala?

A1: He telephoned in order to influence the voting patterns in Dixville Notch.

A2: Because he was concerned that some voters were actually contemplating voting for rival Buchanan.

In this latter example, and in almost all other cases where the answers to a question differ, the two answers do not conflict with each other. Both are possible answers to the question. In this specific case, answer 1 describes the goal that Mr. Bush had with phoning and answer 2 gives Bush's internal motivation for making the phone call.

In many cases, both answers are part of the same explanation, but each of them refers to another item in the reasoning chain. For example,

Q: Why did McDonalds use Paul Bocuse's picture in their advertising campaign?
A1: They wanted to show people in different situations dreaming of Big Macs.
A2: Because they needed a picture of a chef with a white hat.

The text to which these answers refer describes (among other events) the process that leads to McDonald's using a picture of a French chef cook (Mr. Bocuse) in their advertising campaign. First, they decided that they wanted to show people in different situations dreaming of Big Macs. (Answer 1) Second, they wanted one of them to be recognizable as a chef. Thus they needed a picture of a chef with a white hat. (Answer 2).

The fact that 40% of the questions get two different answers supported by the source text leads to the assumption that for many questions, it is not possible to define one single answer as the correct one. This is relevant for the evaluation of a system for *why*-QA. The reference answer should either be a list of possible answers, or the complete reasoning chain (containing these answers) that lies behind the topic of the question.

3.3.1. Ambiguity

As we saw above, many questions can get more than one answer that is supported by the source text. Therefore, it seems that *why*-questions tend to be ambiguous. Particularly interesting in this respect are questions that contain a declarative layer. These questions contain a structural ambiguity that originates from their syntactic form. For example, consider the question below:

Q: Why did Mr. Bocuse say he will use the damages to build a cooking school?

This question can be interpreted in two ways: *Why did Mr. Bocuse say it?* or *Why will Mr. Bocuse use the damages to build a cooking school?* Two participants addressed this question, and both answered according to the last interpretation. However, the similar question

Q: Why does Mr. Jarvis say that the questioning was about as fierce as a spell of underarm bowling with a soft ball?

was answered by two participants according to the first interpretation: *Why does Mr. Jarvis say it?*

This shows that declarative layer questions are structurally ambiguous, which may be relevant for future development of our approach for *why*-QA.

3.3.2. Answer types

As we saw above, *why*-questions can often get more than one defendable answer. For an approach to *why*-QA, it is important to know what kind of items should be looked for in the text. Therefore, we investigated which answer types are possible for *why*-questions.

In earlier work on question classification (e.g. Moldovan et al., 2000), *why*-questions share the single answer type *reason*. Based on the classification of adverbial clauses by Quirk (1985: 15.45), we distinguish the following sub-types of *reason*:

(1) *Cause* which is a temporal relation between two events in which no deliberate human intention is involved. For example,

Q: Why did compilers of the OED have an easier time?
A: Because the OED was compiled in the 19th century when language was not developing as fast as it is today.

(2) *Motivation* which adds a human intention to a temporal causal relation. A motivation can be either a future goal or some person's internal motivation (as we already saw for the two answers to the Bush-question above). For example,

Q: Why has the team of researchers been split up into two teams?
A: To complete the work more quickly – one team will finish "A" while the second team will start on "B".

(3) *Circumstance* which adds conditionality to the temporal relation: the first event is a strict condition for the second event. For example, in the question-answer pair below, the situation described in the answer is a condition for the topic of the question: people will only buy Windows if it works well enough.

Q: Why will people buy Windows?
S: Because it offers more software, it is more fun to use and it works well enough.

(4) *Generic purpose* which does not express a temporal relation between two events, but gives the physical function of an object in the real world. For example,

Q: Why do people have eyebrows?
S: People have eyebrows to prevent sweat running into their eyes.

We manually classified our complete set of question-answer pairs using the four answer sub-types (*cause*, *motivation*, *circumstance* and *purpose*). We assigned the sub-type *circumstance* to a question-answer pair if the answer was a condition for the topic of the question. We assigned the sub-type *motivation* if the answer gave a person's intention for the deliberate action given by the question. If no conditionality or human intention was involved in the temporal relation between question and answer, we assigned the sub-type *cause*. If the relation between question and answer is not temporal, but the answer gives the function of the object given in the question, the answer type *generic purpose* was assigned.

We assigned the answer type *cause* to 52 percent of the questions, *motivation* to 37.5 percent and six question-answer pairs (1.5 percent) were labeled as *circumstance*.

There are no occurrences of question-answer pairs describing *generic purpose* in our data set. This type of relation is very rare in news texts because of its generic character. News texts mainly describe a series of events, states and actions that are specific for the time, place and topic of the text. Generic information like the physical purposes of objects are not commonly given in a text on a specific news topic.

To the remaining pairs (8.9 percent), no answer subtype was assigned. Many of these could not be classified because they do not actually refer to an event in the text, but to some presumed or understood meta-information about the text, such as the reason that a specific topic from the text is relevant or striking, e.g.

```
Q: Why is it surprising that the Supreme
   Court will reopen the abortion debate?
A: Because this means possibly over-ruling
   its 1973 declaration that women have a
   constitutional right to an abortion free
   from government intrusion.
```

The classification of question-answer pairs into answer types is not a straightforward task. We used human intention as a key in distinguishing between cause and motivation, but still the judgments are somewhat subjective. Often, there is a thin line between intentional and unintentional action relations. For example, consider the question below that asks for the reason for Mr. Bocuse's emotions after he received a letter of apology from McDonalds:

```
Q: Why was Bocuse even more angry after the
   letter of apology?
A: Because it became clear that he was not
   well-known in the Netherlands at all,
   which he regarded as an insult.
```

Although the letter of apology itself was a deliberate action, its result was not intended. Therefore, we classified this question-answer pair as *cause*. However, one could also argue that for Mr. Bocuse, hearing that he was not well-known in the Netherlands was his internal *motivation* of being angry.

The difficulty and subjectivity of the classification into answer types raises the question whether knowledge about the answer type is nearly as effective in *why*-QA as it appeared to be in QA for factoids.

4. Conclusions and implications

We have created a data collection for research and development of an approach for *why*-QA. The resulting data set, collected by elicitation, comprises 395 *why*-questions. For each question, the source document and one or two user-formulated answers are available in the data set. For a subset of the questions, the data set also includes user-formulated paraphrases. The question-answer pairs have been annotated with information on topic and answer type. The resulting data set is of importance not only for our research, but we expect it to contribute to and stimulate other research in the field of *why*-QA.[1]

In the reminder of this section, we will discuss the implications of our findings for the development for an approach for *why*-QA.

In the start of section 3, we assumed that our data collection is syntactically and semantically representative for the range of *why*-questions that are formulated in the context of a QA system. This assumption was supported by (1) the similar findings for syntactic variation in our set of questions and the Webclopedia data set (section 3.2.1) and (2) the finding that the differences between question and source text are relatively large (section 3.2.2).

We found (section 3.2.1) that 94% of *why*-questions have the unmarked default word order, which makes syntactic analysis of *why*-questions (step 1 of the QA process) relatively straightforward. On the other hand, the difficulty and subjectivity of the classification into answer types raises the question whether determination of the answer type will contribute substantially to improving the performance of a system for *why*-QA (section 3.3.2)

For matching a question to a source text that is likely to contain the answer (step 2 of the QA process), at least a thesaurus or a semantic net like WordNet is needed for solving the lexical differences. Also, morphological analysis is necessary for matching form variants to one lemma (section 3.2.2).

For step 3 of the QA process, paragraph analysis, it is important to know that many questions have more than one possible answer (section 3.3). Often, the different answers are part of the same reasoning chain. It would help the *why*-QA process if we would succeed in extracting these reasoning chains from the source texts. In the near future, we will investigate the use of Rhetorical Structure Theory (Mann, 1988).

5. References

British National Corpus, 2002. *The BNC Sampler.* Oxford University Computing Services.

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

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

Jijkoun, V. and De Rijke, M., 2005. Retrieving Answers from Frequently Asked Questions Pages on the Web. In: *Proceedings CIKM-2005*, to appear.

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

Mann, W. C., and Thompson, S. A., 1988. Rhetorical Structure Theory: Toward a functional theory of text organization. *Text*, 8(3), 243-281

Maybury , M., 2003. Toward a Question Answering Roadmap. In *New Directions in Question Answering 2003*: 8-11

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

Text Retrieval Conference (TREC) QA track, 2003. http://trec.nist.gov/data/qamain.html

Voorhees, E. and Tice, D., 2000. Building a Question Answering Test Collection. In *Proceedings of SIGIR-2000*: 200-207

Voorhees, E., 2003. Overview of the TREC 2003 Question Answering Track. In *Overview of TREC 2003*: 1-13

[1] The collected data are available on http://lands.let.ru.nl/TSpublic/sverbern/