Creating the DISEQuA Corpus: a Test Set for Multilingual Question Answering

Bernardo Magnini*, Simone Romagnoli*, Alessandro Vallin*

Jesús Herrera**, Anselmo Peñas**, Víctor Peinado**, Felisa Verdejo**

Maarten de Rijke***

* ITC-irst, Centro per la Ricerca Scientifica e Tecnologica
Via Sommarive, 38050 Povo (TN), Italy.
{magnini,romagnoli,vallin}@itc.it

** UNED, Spanish Distance Learning University, Dpto. Lenguajes y Sistemas Informaticos
Ciudad Universitaria, c./Juan del Rosal 16, 28040 Madrid, Spain.
{anselmo,felisa,jesus.herrera,victor}@lsi.uned.es

*** Language and Inference Technology Group, ILLC, University of Amsterdam
Nieuwe Achtergracht 166, 1018 WV Amsterdam, The Netherlands.
mdr@science.uva.nl

Abstract. This paper describes the procedure adopted by the three coordinators of the CLEF 2003 question answering track (ITC-irst, UNED and ILLC) to create the question set for the monolingual tasks. Despite the little resources available, the three groups collaborated and managed to formulate and verify a large pool of original questions posed in three different languages: Dutch, Italian and Spanish. A part of these queries was translated into English and shared between the three coordination groups. Thus, a second cross-verification was conducted, in order to extract the queries that had an answer in all the three monolingual document collections. Finally, the result of the joint efforts was the creation of the DISEQuA (Dutch Italian Spanish English Questions and Answers) corpus, a useful and reusable resource that is freely available for the research community. The article reports on the different stages of the corpus creation, from the monolingual kernels to the multilingual extension.

1 Introduction

The question answering (QA) track at CLEF 2003, starting from the experiences accumulated during the past TREC campaigns, focused on the evaluation of QA systems created for non-English European languages and consequently promoted both monolingual (Dutch, Italian and Spanish) and cross-language tasks. Cross-linguality was a necessary step to push participants into designing systems that can find answers in languages different from the source language of the queries, which mirrors a possible scenario of future applications.

The document collections were those used at CLEF 2002, i.e. articles drawn from newspapers and news agencies of the year 1994 (Dutch, Italian, Spanish) and 1995

(Dutch). Nevertheless, as coordinators of the monolingual tasks, we first needed to create a corpus of questions with related answers for the evaluation exercise, i.e. a replicable gold standard.

According to the CLEF QA guidelines, that are based on last years' TREC ones, the question set released to participants should be made up of simple, mostly short, straightforward and 'factoid' queries. Systems should process questions that sound naturally spontaneous, and a good, realistic question set should consist of questions arisen from a real desire to know something about a particular event or situation. Actually, we could have extracted our questions directly from the document collection, simply turning assertive statements into interrogative ones. Such a procedure would have turned out to be quite quick and pragmatic, but it would have undermined the original intentions of the QA track, which is to evaluate the systems' performance in finding possible answers to open domain questions, independently from the target document collection used. Drawing the queries from the corpus itself would have influenced us in the topics and words choice, and in the syntactic formulation of the questions.

The coordinators of the TREC 2002 QA track obtained their 500 questions corpus from question logs of WWW search engines (like the MSN portal). They extracted a thousand queries that satisfied determined patterns from the millions of questions registered in the logs, and then, after correcting linguistic errors, they searched the answers in a 3GB wide corpus. Similarly, the organizers of the TREC-8 QA (held in 1999) drew one hundred of the 200 final questions from a pool of 1,500 candidate questions contained in the FAQFinder logs [3].

This strategy leads to a well formed questions and answers corpus, but it requires a lot of available resources, i.e. many native speaker people involved in the verification of the questions, a huge document collection, the access to the logs borrowed from search engines companies and - last but not least – a considerable amount of time. We could take advantage neither of question logs nor of a corpus big enough to enable the extraction of any kind of answer. In order to cope with this lack of resources, we conceived an alternative approach to the QA corpus creation, trying to preserve spontaneity of formulation and independence from the documents collection.

The monolingual tasks of the CLEF 2003 QA track required a test set of 200 fact-based questions. Our goal was to collect a heterogeneous set of queries that would represent an extensive range of subjects and find their related answers in three different corpora. The creation of the three test sets constituted the first step toward the generation of a multilingual corpus of questions and answers, whose entries are written into four languages, with the related responses that the assessors extracted from each monolingual document collection during the verification phase.

Our activity could be roughly divided into four steps:

1. Formulation of a pool of 200 candidate questions with their answers in each language;
2. Selection of 150 questions from each monolingual set and their translation into English in order to share them with the other groups;
3. Second translation and further processing of each shared question in two different document collections;
4. Data merging and final construction of the DISEQuA corpus.

2 Question Generation

The corpora addressed by the questions for the monolingual tasks were three collection of newspaper and news agency documents released in 1994 and 1995, and written in Dutch, Italian and Spanish respectively. We used the document collections licensed by the Cross Language Evaluation Forum. These articles constituted a heterogeneous, open domain text collection. Each article had a unique identifier, i.e. a DOCID number, that participants' systems had to return together with the answer string in order to prove that their responses were supported by the text. The text of the Italian collection was constituted by about 27 millions words (200 Mb) drawn from the newspaper *La Stampa* and the Swiss-Italian *SDA* press agency. The Spanish corpus contained more than 200,000 international news from the *EFE* press agency published during the year 1994. The Dutch collection was the CLEF 2002 Dutch collection, which consists of the 1994 and 1995 editions of *Algemeen Dagblad* and *NRC Handelsblad* (about 200,000 documents, or 540 Mb).

```
<DOC>
<DOCNO>EFE19940101-00001</DOCNO>
<DOCID>EFE19940101-00001</DOCID>
<DATE>19940101</DATE>
<TIME>00.28</TIME>
<SCATE>POX</SCATE>
<FICHEROS>94F.JPG</FICHEROS>
<DESTINO>ICX EXG</DESTINO>
<CATEGORY>POLITICA</CATEGORY>
<CLAVE>DP2403</CLAVE>
<NUM>736</NUM>
<TITLE>  GUINEA-OBIANG
        PRESIDENTE SUGIERE RECHAZARA AYUDA EXTERIOR
CONDICIONADA
</TITLE>
<TEXT>  Malabo, 31 dic (EFE).- El presidente de Guinea Ecuatorial, Teodoro Obiang
Nguema, sugirió hoy, viernes, que su Gobierno podría rechazar la ayuda internacional que
recibe si ésta se condiciona a que en el país haya "convulsiones políticas".
   En su discurso de fin de año, [.....] conceptos de libertad, seguridad ciudadana y
desarrollo económico y social. EFE
  DN/FMR
  01/01/00-28/94
</TEXT>
</DOC>
```

Fig. 1. Format of the target document collection (example drawn from the Spanish corpus)

The textual contents of the Spanish collection, as shown in figure 1, were not tagged in any way. The text sections of the Italian corpus on the contrary, according to the NIST guidelines, had been annotated with named entities tags such as <PERSON>, <LOCATION> and <AUTHOR>. The Dutch collections were formatted similarly.

Given these three corpora, our final goal was to formulate a set of 180 fact-based questions shared by all the three monolingual QA tasks. The intention of having the same queries in all the tasks was motivated by the need of comparing the systems' performance in different languages. Since the track was divided in many tasks and

most of the participants took part in just one of them, the use of the same test set, although it was translated into other languages, would allow us to compare the accuracy of different runs. Besides 180 shared queries, we planned to include in each test set 20 questions with no answer in the corpora (the so-called NIL questions).

2.1 From Topics to Keywords

The key element that guided our activity through the first phase of questions generation was the CLEF collection of topics. If we had asked people to generate questions without any restraint, we could have probably obtained just a few usable queries for our purpose. Otherwise, it would have been even more difficult to ask them to focus just on events occurred in 1994 or 1995, which is the time coverage of the articles in our text collections. Besides, we noticed that the mental process of conceiving fact-based questions without having any topic details could take a considerable amount of time: asking good questions could be as difficult as giving consistent answers. In order to cope with these drawbacks, to improve the relevance of the queries and to reduce the time necessary to their generation, we decided to use some CLEF topics.

Topics, that can be defined as "original user requests" [1], represent a resource developed for many NLP applications, included question answering. The team that generated the CLEF topics wanted to create a set of real life subjects which should meet the contents of the document collections. The main international political, social, cultural, economic, scientific and sporting issues and events occurred in 1994 and 1995 were included and topics were written in a SGML style, with three textual fields, as in figure 2.

<top>
<num> C001
<I-title>
Architettura a Berlino
<I-desc>
Trova documenti che riguardano l'architettura a Berlino.
<I-narr>
I documenti rilevanti parlano, in generale, degli aspetti architettonici di Berlino o, in particolare, della ricostruzione di alcuni parti della città dopo la caduta del Muro.
</top>[1]

Fig. 2. An Italian topic released by CLEF in the year 2000 (translation in the footnote)

The title field sketches straightforwardly the main content of the topic, the description field mirrors the needs of a potential user, presenting a more precise formulation in one sentence, and the narrative field gives more information concerning relevance.

[1] <I-title>Architecture in Berlin
<I-desc>Retrieve documents that concern architecture in Berlin.
<I-narr>Generally speaking, the relevant documents deal with the architectural features of Berlin or, particularly, with the reconstruction of some parts of the city after the knocking down of the Wall.

In the very first experiment ITC-irst carried out to generate its questions set, two volunteers were provided with three CLEF topics structured as above, asking them to produce ten queries for each one. It took about forty-five minutes to conclude their task, and it was immediately noticed that the questions were too closely related to the source topics. Therefore this pilot experiment showed the weaknesses and drawbacks of the strategy, which would lead to overspecified questions, and underlined the need to improve the stimulating power of the topics reducing their specificity without losing relevance to the corpus.
The simplest way to expand the structure of the topics and widen the scope of activity for the people in charge of the questions generation seemed to extract manually from each topic a series of relevant keywords, that would replace the topics themselves. No particularly detailed instructions were given in that phase: we just isolated the most semantically relevant words. A keyword could be defined as an independent, unambiguous and precise element that is meant to arise interest and stimulate questions over a specific issue. We also inferred keywords that were not explicitly present in the topic, assuming that even external knowledge, though related to the topic, could help to formulate pertinent questions. ITC-irst coordinators took into consideration the topics developed by CLEF in the years 2000, 2001 and 2002. Three people were involved in the extraction of keywords, that were appended to each topic in form of a 'signature', as the tag in the following example testifies. So, the topic entitled "Architecture in Berlin" (shown in figure 2) was converted into a list of word that could even appear unrelated to each other:

<IT-tsig>
architettura, Berlino, documenti, aspetti architettonici, ricostruzione, città, caduta del Muro, Muro
</IT-tsig>[2]

It is interesting to notice that the keywords, even though originated from the topics, allowed a certain detachment from the restricted coverage of the topics themselves, without losing the relation with the important issues of the years 1994 and 1995, that constituted the core of the document collection. Thus the experiment was repeated and much better results in terms of variety and generality of the queries were achieved, in fact the people who were given the keywords instead of the topics had more freedom to range over a series of concepts without any restraint or conditions of adherence to a single specific and detailed issue. Though the nearness of correlated keywords led to the generation of similar queries, this strategy was definitely adopted.

The CLEF topics had a pivotal role also in the generation of the Spanish and Dutch queries. As a preparatory work, the Spanish UNED NLP group studied the test set used at TREC 2002 and tried to draw some conclusions in terms of the questions formulation style and the necessary casuistry to find the answer. Then, four people were given the CLEF topics of the years 2000, 2001 and 2002 (but no keywords) with the task of producing 200 short, fact-based queries. The Dutch LIT group adopted the same strategy in its preparation. TREC QA topics (1-1893) were translated into Dutch, and old CLEF retrieval topics (1-140) were used to generate Dutch example questions, usually around 3 per topic.

[2] Architecture, Berlin , documents, architectural aspects, reconstruction, city, knocking down of the Wall, Wall.

2.2 From Keywords to Questions

Before generating the queries, the three groups agreed on common guidelines that would help to formulate a good and useful test set. Following the model of past TREC campaigns, and particularly of the TREC 2002 QA track, a series of basic instructions were formulated.

Firstly, questions should be fact-based, and, if possible, they should address events that occurred in the years 1994 or 1995. When a precise reference to these two years lacked in the questions, it had to be considered that systems would use a document collection of that year. No particular restraints were imposed on the length and on the syntactic form of the queries, but coordinators kept them simple.

Secondly, questions should ask for an entity (i.e. a person, a location, a date, a measure or a concrete object), avoiding subjective opinions or explanations. So, "Why-questions" were not allowed. Queries like "Why does Bush want to attack Iraq?" or "Who is the most important Italian politician of the twentieth century?" could not be accepted.

Since the TREC 2002 question set constituted a good term of comparison, and it did not include any definition question of the form "Who/What is X?", it was decided to avoid them, as well.

Thirdly, coordinators agreed that multiple-item questions, like those used in the TREC list-task, should be avoided. If the community will be interested in processing list questions, we could propose them in next year's track, possibly together with definition queries. As a pilot evaluation exercise, we did not want to introduce too many difficulties that could have discouraged potential participants.

Similarly, the people in charge for the questions generation could not formulate 'double queries', in which there is a second indirect question subsumed within the main one (for instance, "Who is the president of the poorest country in the world?").

Finally, closed questions, known as yes/no questions, should be left out, too. Queries should be related to the topics or to the keywords extracted from the topics, without any particular restraint in the word choice. It was not necessary to know the answer before formulating a question: on the contrary, assessors had to be as close as possible to the information they found in the document collection. A prior knowledge of the answer could influence the search in the corpus.

Given these instructions, thirty people at ITC-irst were provided with two sets of keywords (extracted from two topics) and were asked to generate ten questions for each one. In this way, a large pool of 600 candidate queries was created. The examples shown in figure 3 demonstrate that the keywords extended the limited scope of the topic "Architecture in Berlin", allowing people to pose questions related to history or even politics. Some questions, as number 5 and 9, lost connection with the original form of the topic, introducing the name of a famous architect and asking for the number of inhabitants rather than focusing on the architectural features of the city. Adopting this strategy, we could preserve a certain adherence to the original content of the topic, introducing some new elements. Inevitably, as a side effect a number of queries turned out to be useless because they were almost unrelated to the keywords or badly formulated.

<num>C001</num>
<keyword> architettura, Berlino, documenti, aspetti architettonici, ricostruzione, città, caduta del Muro, Muro </keyword>
<question n=1> Quando e' caduto il muro di Berlino? </question>
<question n=2> Chi ha costruito il Muro di Berlino? </question>
<question n=3> Quanto era lungo il muro di Berlino? </question>
<question n=4> Qual e' la piazza piu' importante di Berlino? </question>
<question n=5> Qual e' la professione di Renzo Piano? </question>
<question n=6> Quando e' stato costruito il muro di Berlino? </question>
<question n=7> Quando e' che Berlino e' ritornata ad essere capitale?</question>
<question n=8> Dove si trova Berlino? </question>
<question n=9> Quanti abitanti ha Berlino? </question>
<question n=10> Che cosa divideva il muro di Berlino? </question> 3

Fig. 3. Questions generated from a list of keywords (translation in the footnote)

In spite of the generation guidelines established before producing the candidate questions, some inconsistencies persisted. For instance, question 4 concerns a personal opinion rather than a fact-based datum: it is not clear how the importance of a place could be objectively measured. Similarly, question 7 deals with events occurred later than 1994: although the German government took the decision in 1991, Berlin officially became the capital city in 1999.

2.3 Questions Verification

Once the candidate questions had been collected, it was necessary to verify whether they had an answer in the target document collection. This phase constituted the actual manual construction of the replicable gold standard for the CLEF QA track: systems would later process the questions automatically.

ITC-irst involved three native Italian speakers in this work. In order to cope with the large amount of candidate questions and with the possibility that many of them were not compliant with the generation guidelines and could not be used for the QA track, three different categories of queries were arranged and each question was classified: the entries of list A were queries that respected the generation guidelines and whose answer was intuitively known, in list B were placed the relevant questions that in the assessors' opinion had a more difficult answer, while list C contained those that were badly formulated or did not respect the guidelines instructions. As expected, list B was the largest one, including 354 questions. At the end of the question

[3] <question n=1> When did the Berlin Wall fall? </question>
<question n=2> Who built the Berlin Wall? </question>
<question n=3> How long was the Berlin Wall? </question>
<question n=4> Which is the most important square in Berlin? </question>
<question n=5> What is Renzo Piano's job? </question>
<question n=6> When was the Berlin Wall built? </question>
<question n=7> When did Berlin become the capital again? </question>
<question n=8> Where is Berlin? </question>
<question n=9> How many inhabitants are there in Berlin? </question>
<question n=10> What did the Berlin Wall divide? </question

verification phase, a total of 480 questions were processed manually, and the remaining 120, most of those included in list C, were eliminated.

Browsing a document collection in search of the answers could be a very exhausting activity without any tool that facilitates the detection of the relevant strings. Fortunately, ITC-irst had available a concordancer[4] that allowed the three assessors to make selective searches within the corpus, to find the correct answers and to go back to the docid, i.e. the unique identifier, of the document that supported each answer. The common strategy employed by the assessors was to type parts of the query or parts of the known answer in the concordancer, and then browse the most relevant documents retrieved by the software in search of a text snippet that justified and supported the correct answer. The Dutch group developed a small number of grep-based shell scripts with the same purpose.

The problem of structuring data and find a sensible format to describe both questions and answers arose during this first phase of the creation of DISEQuA. The issue was addressed conceiving an XML syntax that would show the number of each question, the keywords set (or topic) from which it was generated, the person who verified it in the document collection and the type of entity it was related to. Similarly, the answers found for each question needed to be numbered, and the docid of the document that supported each response had to be logged. The adoption of a precise format could solve the problem of losing trace of the changes that each question could undergo, in fact new tags could be added to give more information. Secondly, structured data can be easily browsed and analyzed: for instance, the tag used to indicate the question type proved to be quite useful in balancing the test set. Thirdly, a common format for questions and answers was necessary to share them between the three groups that put together the DISEQuA corpus.

Figure 4 shows an example drawn from the Italian question set : the attribute ‘cnt’ indicates the number assigned to the question, ‘assessor’ is an identifier of the person who processed the query, which seemed to be important in case of inconsistencies. In the attribute ‘origin’ is given the name of the file containing the keywords extracted from a single topic, while the attribute ‘type’ describes the category to which the answer belongs. Seven different question types were considered: PERSON, LOCATION, MEASURE, DATE, ORGANIZATION, OBJECT (i.e. concrete things) and OTHER (when the response could not be labeled with one precise type). The aim was to create a well-balanced test set, with a good coverage of all these categories.

Likewise, the attribute ‘n’ in the tag <answer> represents a progressive number of responses, in fact a single query could have several correct answers in the same document collection. Dates and numbers in particular change across different news for the same event. Sometimes former news in the document collection are less precise than the latter ones, because they register a process that changes over a period. Since systems were expected to give an answer supported by a unique document, and not the final or best answer in the whole corpus, in such cases there were many correct responses. In the attribute ‘idx’ is given the docid identifier of the document in which each single answer appears. Systems should return the docid as a justification of the answer, and in strict evaluation the unsupported responses were considered as incorrect.

[4] the “Toolbox for Lexicographers” developed by Claudio Giuliano.

```
<qa>
    <question cnt="42" assessor="ALE" origin="keyword_C001.txt" type="MEASURE">
            Quanti abitanti ha Berlino?
    </question>
    <answer n="1" idx="SDA19940804.00147">
             3,5 milioni
    </answer>
</qa>
```

Fig. 4. Format of the verified questions (see question 9 in figure 3)

When no answer was found in the target corpus, answer 'n' and 'idx' were labeled with 0 (zero), and the answer string was replaced by the string "NIL". Queries with no answer were not eliminated: on the contrary, twenty NIL questions were included in the final version of each monolingual test set to evaluate systems' accuracy in recognizing that there was no response.
Sometimes the responsiveness of the retrieved string was doubtful and the assessors could not decide whether it was acceptable. These cases required a deeper analyses and an agreement between different assessors. In order to signal the doubts that emerged during the verification phase, a "star" character (*) was put before the uncertain answers and a significant remark that justified the uncertainty was appended to the question within the tag <rem>, as in the following example (see question 10 in figure 3):

```
<question n=5 origin=keyword_C001 type=LOCATION>
  Che cosa divideva il muro di Berlino ?
</question>
*<answer n="1"  idx="LASTAMPA19941016.00038">
  Germania
</answer>
*<answer n="2"  idx="LASTAMPA19941016.00038">
  mondo
</answer>
<rem>"Un evento inatteso, spettacolare, emozionante: sotto gli occhi del mondo cade il Muro di
Berlino, simbolomateriale della divisione della Germania e del mondo."</rem>5
```

A cut-and-pasted text snippet found in the document collection was usually placed in the tag <rem>, so that another assessor could take a decision without opening again the corpus in search of the necessary contextual information. In the example above, it was not clear whether the retrieved answers, which are metaphorical, could be accepted (actually, the Berlin Wall isolated West Berlin from the German Democratic Republic), so the first assessor that processed the question left the response undetermined. If a second assessor could not take a decision, the question was passed to a third person, who normally solved the doubts. Alternatively, badly-formulated questions could be slightly modified in order to match the retrieved answer.

---

[5] *<answer n="1"> Germany
*<answer n="2"> the world
<rem>"An unexpected, spectacular and exciting event: the eyes of the world are on the Berlin Wall that is falling, a concrete symbol of Germany's and the world's division."</rem>

Some candidate questions asked for events occurring “in the year 1994” (or 1995), but since 1994 (and, for Dutch, 1995) was the year in which the target corpora were published, it was very improbable that it would appear explicitly in the articles, so no document would clearly state that the year was 1994 (or 1995). For this reason, every explicit mention of the year 1994 (or 1995) had to be removed from the final version of the queries.

3 Questions Sharing

At this point, each group had collected and verified 200 questions formulated in its own language. A small part of them (10%) had no answer in the document collections, while the other ones had at least one supported response. Since the aim was to create a multilingual test set whose entries were formulated into three different languages, it was necessary to share the questions that had been generated independently. Thus, each group selected the 150 queries that seemed most likely to find an answer also in the other two document collections and translated them into English before sending them to the larger pool. Figure 5 shows the format chosen for the questions sharing. The questions that were too strictly related to the issues or events of a particular country were skipped.

```
<qa cnt="20" type="MEASURE">
  <language val="ITA" original="TRUE">
        <question assessor="ALE">
              Quanti abitanti ha Berlino?
        </question>
        <answer n="1" idx="SDA19940804.00147">
              3,5 milioni
        </answer>
  </language>
  <language val="ENG" original="FALSE">
        <question assessor="">
              How many inhabitants are there in Berlin?
        </question>
        <answer n="1" idx="-1">
              SEARCH[3,500,000]
        </answer>
  </language>
</qa>
```

Fig. 5. Question sharing format

English was chosen as intermediate language for two reasons: firstly to build a richer linguistic resource for further QA evaluation, considering that most question answering systems are currently designed for English applications; secondly, to simplify the passage from one language to another, without recurring to professional translators. Nevertheless, the translation into English required much attention because in the passage from one language to another, the syntactic formulation or even the meaning could change. It is important to underline that the 450 questions that form the DISEQuA corpus underwent three translations: one from the source language into English and then other two from English into the two target languages. Each

translation could introduce some variations, with the risk that the four final versions would not be semantically equivalent and aligned. To avoid this problem, in the second translation both the English version and the original question in the source language were taken into consideration.
If we compare figures 4 and 5, we see that important changes in the format were introduced in this phase. Though the question is the same in the two figures, it is numbered differently, in fact some questions that were placed before this one in the monolingual Italian test set were discarded because they had little chances of finding an answer in the Dutch and Spanish corpora.
The new tag <language> was added, with its attributes 'val' and 'original'. The former indicates the language in which the question appears ("DUT", "ITA", "SPA" or "ENG"). The latter keeps track of the source and the target language of each query: 'original' can have either "TRUE" or "FALSE" as Boolean values, where "TRUE" shows that the 'language val' is the source language, i.e. the language in which the question was first generated, while "FALSE" records that the query has been translated. Consequently, English questions, as intermediate versions, could have nothing but "FALSE".
Concerning the answer string format, a default negative value "-1" was assigned to the English version of each question, to distinguish it from the zero used in NIL questions. The string "SEARCH" followed by the translation within square brackets of the correct answer found in the source corpus constituted a valuable help for the assessors who would process the shared questions.

4 Data Merging

Summarizing, each group selected and translated 150 verified questions from its monolingual test set, so that a large pool of 450 queries formulated into English and a second source language was created. In the following phase, each group picked up the 300 questions submitted by the other two and translated them a second time from English into a new target language. As a consequence, all the questions had a translation in four different languages and could be processed again in the other two target document collections.
When the second verification was concluded, the resulting data were merged. The different versions of the same questions were aligned, and the DISEQuA corpus was successfully assembled. Figure 6 shows how each question appears in the multilingual test set.
The merging revealed that 246 questions had at least one answer in all the three reference document collections, 111 had at least a response in two of them, and the remaining 93 just in the source corpus in which they were first processed. A subset of 180 shared questions with answer in all the three corpora was randomly extracted from the merged collection. Each group then added 20 NIL questions in order to create its final monolingual test set for the CLEF 2003 QA track. Due to lack of time, the 180 queries that the three test sets had in common could not be chosen manually and attentively, but fortunately they turned out to be quite balanced: 45 entries asked for the name or role of a PERSON, 40 pertained a LOCATION, 31 a MEASURE, 23

an ORGANISATION, 19 a DATE, 9 a concrete OBJECT, and 13 could be labeled with OTHER.
During the merging, it was noticed that some questions had the same meaning: 13 duplicates were found, but since most of them were formulated in a slightly different way, which did not affect the semantic contents, it was decided to keep them in the DISEQuA corpus. Different formulations of the same question could be exploited in Machine Translation applications.

4.1 Availability of the DISEQuA Corpus

The DISEQuA corpus is the result of the joint effort of three research groups; the effort aimed at creating not only a good test set for the CLEF QA track, but also a useful and reusable resource for further QA evaluation. It is freely available on the "QA @ CLEF" web site[6], together with the question set developed for the bilingual tasks of the CLEF competition. Both can be used for NLP applications, and everyone can download them and introduce further material, adding questions, answers or even queries in other languages.
Together with the DISEQuA corpus, another test set is available: a collection of 200 English questions translated into Dutch, French, German, Italian and Spanish. It is the test set created for the CLEF QA cross-language tasks. Differently from DISEQuA, only one target corpus was used to verify the queries of this second resource. So, each question has six different translations, but the answers have been searched only in the *Los Angeles Times* document collection.

5 Conclusions

In this paper we outlined the procedure used for creating a multilingual corpus of questions and answers. The resource we developed constitutes a reusable source of information for many NLP fields. We translated 450 questions into four languages (Dutch, Italian, Spanish and English) and processed them in three different target corpora, retrieving the answers and the docid of the documents that support the answers. So the queries we generated can be employed to evaluate translingual QA in 12 different combinations.
In the future we could search for answers in an English document collection, for instance the Los Angeles Times corpus licensed by CLEF, and widen the scope of possible applications. DISEQuA could be updated by adding other questions in different languages or other target corpora. The focus of new queries could not be limited to simple factoid questions, but it could address definitions or lists of items, as well. The tiny corpus we built could be enriched in several ways.
It can be used also in Machine Translation, because questions have particular features that other corpora do not usually address and that deserve to be investigated. Any further development will constitute an enrichment of the resource.

[6] http://clef-qa.itc.it

```
<qa cnt="20" type="MEASURE">
        <language val="ITA" original="TRUE">
            <question assessor="Ale-irst">
                    Quanti abitanti ha Berlino?
            </question>
            <answer n="1" idx="SDA19940804.00147">
                    3,5 milioni
            </answer>
        </language>
        <language val="SPA" original="FALSE">
                <question assessor="Víctor-UNED">
                        ¿Cuántos  habitantes tiene Berlín?
                </question>
                <answer n="1" idx="EFE19940107-02622">
                        Casi cuatro millones
                </answer>
        </language>
        <language val="DUT" original="FALSE">
                <question assessor="LIT">
                        Hoeveel inwoners heeft Berlijn?
                </question>
                <answer n="1" idx="NH19950601-0163">
                        3,5 miljoen
                </answer>
        </language>
        <language val="ENG" original="FALSE">
                <question assessor="">
                        How many inhabitants are there in Berlin?
                </question>
                <answer n="1" idx="-1">
                        SEARCH[3,500,000]
                </answer>
        </language>
</qa>
```

Fig. 6. Final question format in the DISEQuA corpus

6 Acknowledgements

The work described in this paper has been supported by the Autonomous Province of Trento, in the framework of the WebFAQ Project, by the Spanish Government (MCyT, TIC-2002-10597-E) and by the Netherlands Organization for Scientific Research (NWO) under project numbers 612-13-001, 365-20-005, 612.069.006, 612.000.106, 220-80-001, 612.000.207, and 612.066.302.
We would like to thank all the people at ITC-irst (TCC division) who posed the necessary questions for the monolingual Italian test set. We are grateful to Claudio Giuliano, who made his “Tool for Lexicographers” available, facilitating the search for answers in the corpora. We are also indebted to Franca Rossi and Elisabetta Fauri at ITC-irst, for their help in the verification of the questions. We also want to thank Henry Chinaski, Vera Hollink, and Valentin Jijkoun for their help in developing and assessing the questions for the monolingual Dutch task. Without their contributions,

the DISEQuA corpus would not exist. Finally, we wish to thank Charles Callaway who, as English native speaker, edited the translation of the Italian questions.

References

1. Kluck M. and Womser-Hacker C.: Inside the Evaluation Process of the Cross Language Evaluation Forum (CLEF): Issues of Multilingual Topic Creation and Multilingual Relevance Assessment. Proceedings of the Third International Conference on Language Resources and Evaluation, (LREC 2002), Las Palmas de Gran Canaria 29-31 May 2002.
2. Magnini B.: Evaluation of Cross-Language Question Answering Systems, proposal presentation held at the CLEF Workshop 2002.
   URL: http://clef.iei.pi.cnr.it:2002/workshop2002/presentations/q-a.pdf.
3. Tice D. M. and Voorhees E. M.: The TREC-8 Question Answering Track Evaluation. Proceedings of the Eighth Text REtrieval Conference (TREC-8), Gaithersburg, MD., 2000.