GUIDELINES for the PARTICIPANTS in the ResPubliQA 2010 exercise

I. TASK OVERVIEW

The aim of ResPubliQA 2010 is to capitalize on what has been achieved in the previous evaluation campaign while at the same time adding a number of refinements:

- The addition of new question types and the refinement of old ones;
- The opportunity to return both paragraph and exact answer;
- The addition of a new document collection: EUROPARL.

Two separate tasks are proposed for the ResPubliQA 2010 evaluation campaign:

1. PARAGRAPH SELECTION (PS) TASK: to retrieve one paragraph (Text+ID) containing the answer to a question in natural language. This task is very similar to the one performed last year.

2. ANSWER SELECTION (AS) TASK: beyond retrieving a paragraph, systems are required to retrieve also the exact answer (shorter string of text) answering a question in natural language.

The two tasks are different only in the output required. Document collections and test data for both tasks are the same.

LANGUAGES INVOLVED: parallel-aligned documents are available in 9 languages, i.e: Bulgarian, Dutch, English, French, German, Italian, Portuguese, Romanian and Spanish.

Cross-lingual tasks: each group is allowed to participate in one or both tasks considering questions and target collections in any pair of languages taken from these nine.

Monolingual tasks: in addition, all nine monolingual tasks will be enacted, including monolingual English.

N.B.: Although all the above languages are available, only the cross-lingual and monolingual sub-tasks for which at least two participants have registered will be activated.

II. IMPORTANT DATES

ResPubliQA 2010: Track Schedule	
from February 1 to March 30	Registration* at the ResPubliQA website http://celct.isti.cnr.it/ResPubliQA/index.php
May 17	Release of test questions
May 27	Hard deadline for participants' submissions
June 25	Release of individual evaluated results
July 10	Submission of papers
July 30	Notification of acceptance
August 10	Submission of camera ready papers
September 20-23	CLEF Workshop (Padua, Italy)

Test questions will be posted on the ResPubliQA website http://celct.isti.cnr.it/ResPubliQA/ on May 17 and submissions will be due within **5 days** from the first test set download and not later than May 27 by 11:59 p.m. (CEST) Late submissions will not be considered.

Participant results will be submitted using an automatic submission procedure. Details about the submission procedure will be provided when the test data is released. Before completing the submission, a checking routine will be automatically run in order to detect format inconsistencies and common errors in the files (invalid document numbers, wrong formats, missing data, etc..). The automatic submission procedure will reject any run which is not compliant with the required format.

LAB PUBLICATIONS

The proceedings of the ResPubliQA workshop will be reviewed by a peer review process and will be assigned a ISBN number. Details about paper submission will be given later.
The proceedings of all LABs will be prepared in digital form only and will be posted on the CLEF website shortly before the workshop.

* To register for participation, the form at the ResPubliQA website must be filled in. The following information is required:

- full names, affiliations, and e-mail addresses of the group,
- the name of the contact person,
- the evaluation task(s) you will be participating in (PS; AS; or both)
- the monolingual languages or cross-lingual language pairs you are interested in, in order to activate the cross-language sub-tasks.

Remember that in order to activate the cross-lingual sub-tasks you have to register at the ResPubliQA website.

You will receive notification of the registration.

III. DOCUMENT COLLECTION

The ResPubliQA collection is made up of a subset of two multilingual parallel aligned document collections:

The JRC-ACQUIS Multilingual Parallel Corpus[1]: is a freely available parallel corpus containing the total body of European Union (EU) documents, of mostly legal nature. It comprises the contents, principles and political objectives of the EU treaties; EU legislation; declarations and resolutions, international agreements; acts and common objectives. Texts cover various subject domains, including economy, health, information technology, law, agriculture, food, politics and more. This collection of legislative text currently comprises selected texts written between 1950 and 2006 with parallel translations in 22 languages.

The corpus is encoded in XML, according to the TEI guidelines.

A sub-set of the JRC-ACQUIS has been created with parallel and aligned documents in all the 9 languages involved in the track (Bulgarian, Dutch, English, French, German, Italian, Portuguese, Romanian and Spanish). The sub-set consists of roughly 10.700 parallel and aligned documents per language[2].

[1] http://wt.jrc.it/lt/Acquis/

[2] Please note that it cannot be guaranteed that a document available on-line exactly reproduces an officially adopted text. Only European Union legislation published in paper editions of the Official Journal of the European Union is deemed authentic.

The Europarl collection[3]: is a collection of the Proceedings of the European Parliament dating back to 1996. It comprises translations of each of the 11 official languages of the European Union (Danish, German, Greek, English, Spanish, Finnish, French, Italian, Dutch, Portuguese and Swedish). With the enlargement of the European Union to 25 member countries in May 2004, the European Union has began to translate texts into even more languages. Translations into Bulgarian and Romanian are present starting from January 2009.

A (very small) subset of the Europarl has been created with parallel documents in all the 9 languages involved in the track (Bulgarian, Dutch, English, French, German, Italian, Portuguese, Romanian and Spanish) by crawling the web to get the data from the European Parliament's website[4]. The sub-set consists of roughly 150 parallel and aligned documents per language, including:

- Debates (CRE) crawled starting from 01/01/2009
- Texts Adopted (TA) crawled starting from 01/01/2007.

N.B.: In order to facilitate the identification of the paragraphs, each of them has been given a unique progressive ID number inside each document.

The subject of the Acquis documents is European legislation while EUROPARL deals with the parliamentary domain. The two collections are different in style and content while being fully compatible at the same time.

Participants can download both collections from the ResPubliQA website
http://celct.isti.cnr.it/ResPubliQA/Downloads

III. I FORMAT of the COLLECTIONS

JRC-ACQUIS:

- documents are sorted according to the language, and inside each language, documents are grouped by year
- all documents have a numerical identifier called the CELEX code (it helps to find the same text in the various languages)
- each document contains the *header* (giving for instance the download URL and the EUROVOC codes) and the *text* (which consists of the title and a series of paragraphs)

[3] http://www.europarl.europa.eu/

[4] The European Parliament website states: "Reproduction of textual data and multimedia items which are the property of the European Parliament (© European Parliament, year) or of third parties (© External source, year) and for which the European Parliament holds the rights of use is authorized for non-commercial purposes only provided that the entire item is reproduced and the source is acknowledged."

- Text is divided into:
  - *Body* text, each paragraph marked with "<p>" tag,[5]
  - Optional *signature* (list of persons names, and references to other documents) and
  - *Annex* (list of addresses, list of goods, etc.).

EUROPARL:

- documents are grouped by language
- each files contains the relevant information for identification, such as its language and the date, for example: EP_CRE-20090112-IT_clean.xml; or EP_TA-20090113-IT_clean.xml. CRE stands for "debate" and TA stands for "Texts Adopted"
- files are xml encoded
- each file is segmented in chapters and each chapter is made up of one or more speeches given by one or more people. The identity of the speaker is clearly identifiable, as it is given in the attribute of the tag <speaker>. The information about the speaker does not have to be processed
- each speaker's talk is divided in different paragraphs marked as <p>. Paragraph length consists typically of 2-5 sentences
- in order to facilitate the identification of the paragraphs, each of them has been given a unique progressive ID number inside each document.

IV. QUESTIONS

The test set will include a pool of 200 questions (only in the languages of the tasks which have been activated, i.e. the tasks for which at least two have registered!).

Questions in the test set will NOT be grouped in series and will NOT contain anaphoric links to other entities.

Each question in the set will be phrased so that the complete answer to it is **contained in one paragraph**. Questions will be of the following types:

a) Factoid
b) Definition
c) Reason-Purpose
d) Procedure

5 Each paragraph of the text collection can be uniquely identified using the language, the CELEX identifier and the paragraph number.

e) Opinion
f) Other

No NIL questions will be provided, and in the case of LIST questions, all the requested items must be included in the response to be returned: a paragraph in the case of the PS task and a short contiguous string of text in the AS task.

a) **Factoid questions** are fact-based questions, asking for the name of a person, a location, the extent of something, the day on which something happened, etc.

Examples:

> **Q:** *What currencies are used for monetary amounts in the Oil Bulletin?*
> **P:** The information forwarded will be published by the Commission in the Oil Bulletin in dollars and in euros. The monthly rate for the euro vis-à-vis the dollar will be established in accordance with the official market exchange rate(1).
> **A:** dollars and in euros
>
> **Q:** *In how many languages is the Official Journal of the Community published?*
> **P:** The Official Journal of the Community shall be published in the four official languages.
> **A:** four official languages

b) **Definition questions** are questions such as "What/Who is X?", questions asking for the role/job/important information about someone, questions asking for the mission/full name/important information about an organization.

Examples:

> **Q:** *What does IPP denote in the context of the environmental policies?*
> **P:** Since then, new policy approaches on sustainable goods and services have been developed. These endeavours undertaken at all political levels have culminated in the Green Paper on Integrated Product Policy(1) (IPP). This document proposes a new strategy to strengthen and refocus product-related environmental policies and develop the market for greener products, which will also be one of the key innovative elements of the sixth environmental action programme - Environment 2010: %quot%Our future, our choice%quot%(2).
> **A:** Integrated Product Policy
>
> **Q:** *What is an area of departure?*

P: Area of departure and area of destination mean respectively the place where the journey begins and the place where the journey ends, together with, in each case, localities within a radius of 50 kilometres.
A: the place where the journey begins

c) **Reason-purpose questions:** questions asking for the reasons/goals for something. happening

Examples:

Q: *Why have imports of live poultry from Romania been suspended?*
P: (2) Commission Decision 2005/710/EC of 13 October 2005 concerning certain protection measures in relation to highly pathogenic avian influenza in Romania [5] provides that Member States are to suspend imports of live poultry, ratites and farmed and wild feathered game and hatching eggs of those species from the whole territory of Romania and of certain products from birds from parts of that territory.
A: highly pathogenic avian influenza in Romania

Q: *What is the Commission's aim in establishing financial forecasts each year, broken down by category of expenditure, for the three subsequent financial years*?
P: In order to place the budget of the Communities within a framework of forward planning for several years, the Commission shall, each year, after receiving the Opinion of the Budgetary Policy Committee, draw up a financial forecast for the three subsequent financial years, showing the financial implications for the Community resulting from Regulations and Decisions in force and from proposals submitted by the Commission to the Council. The forecast shall be broken down by category of expenditure.
A: to place the budget of the Communities within a framework of forward planning for several years

d) **Procedure questions:** questions asking for a set of actions which is the official or accepted way of doing something:

Examples:

Q: *How do you find the maximum speed of a vehicle?*
P: The maximum speed of the vehicle is expressed in kilometres per hour by the figure corresponding to the closest whole number to the arithmetical mean of the values for the speeds measured during the two consecutive tests, which must not diverge by more than 3 %. When this arithmetical mean lies exactly between two whole members it is rounded up to the next highest number.

A: the figure corresponding to the closest whole number to the arithmetical mean of the values for the speeds measured during the two consecutive tests, which must not diverge by more than 3 %. When this arithmetical mean lies exactly between two whole members it is rounded up to the next highest number.

Q: *How do you measure the stopping distance of a tractor?*
P: 1. BRAKING TESTS 1.1. General 1.1.1. The performance prescribed for service braking devices shall be based on the mean deceleration calculated over the stopping distance. The stopping distance shall be the distance covered by the tractor from the moment when the driver begins to actuate the control of the device until the moment when the tractor stops.
A: The stopping distance shall be the distance covered by the tractor from the moment when the driver begins to actuate the control of the device until the moment when the tractor stops.

e) Opinion questions: question asking for the opinions/feelings/ideas about people, topics, events…

Examples:

Q: *What is the European Parliament position with respect to terrorism?*
P: Today, we in the European Parliament would like to speak out loudly and clearly against the indiscriminate violence of terrorism. We condemn utterly the senseless destruction of human life, the deaths of entire families as a result of blind fanaticism, which causes people to kill their fellow human beings and to trample human dignity underfoot. Terrorism is a direct attack on freedom, human rights and democracy. Terrorism is an attempt to destroy by means of indiscriminate violence the values which unite us in the European Union and within our Member States.
A: Today, we in the European Parliament would like to speak out loudly and clearly against the indiscriminate violence of terrorism.

Q: *What is the opinion of the Committee on Employment and Social Affairs about social cohesion?*
P: Mr President, Mr Barroso, it is the unanimous opinion of the Committee on Employment and Social Affairs that I should like to share with you this morning, for we are looking to see a real promotion of social cohesion in this recovery plan. Social cohesion means being integrated into the labour market. To begin with, then, we want to keep all employees in their jobs and get the unemployed back to work by, among other things, directing the Globalisation Adjustment Fund towards new training courses, so that the workforce is prepared for when we emerge from the crisis.

A: we are looking to see a real promotion of social cohesion in this recovery plan.

f) 'Other' questions: A small number of experimental questions not falling into the previous categories will be included in the set.

V. RESPONSES

Each submitted run must contain an answer (or NOA for NO Answer) for each question. Partial submissions will not be accepted.

Each question must receive one of the following responses depending on the selected task.

PARAGRAPH SELECTION TASK (PS):

either a) The paragraph containing the candidate answer. All the information about the answer must be contained in the paragraph to be returned. Paragraphs to be returned are explicitly marked in the documents of both collections by the mark <p> and their corresponding identifiers.

or b) The string NOA to indicate that the system prefers not to answer the question (see next section).

Each paragraph returned by the system is required to be an extract from a document in the parallel corpus, so the passage string must be simply cut and pasted from the corresponding document. No automatic generation of information contained in different sections of the relevant document is allowed.

Each paragraph is supposed to contain the answer to the question. The selected paragraph must provide enough context to make it clear for the manual assessors if the answer is responsive or not.

There is no maximum length of paragraph retrieved but the **entire <p>** should be returned; even in case of long paragraphs, they should not be truncated.

Each paragraph returned must be supported by:

1. The associated document [*doc_id*] of the ACQUIS or EUROPARL corpora from which the paragraph has been extracted;
2. The id of the paragraph [*p_id*] from which the paragraph has been cut/pasted.

ANSWER SELECTION TASK (AS):

either a) The paragraph containing the candidate answer together with a shorter string of text corresponding to the exact answer.

or b) The string NOA to indicate that the system prefers not to answer the question.

Each exact answer returned is a *continuous* portion of text which answers the question: it must be cut/pasted from the paragraph from which it has been extracted. There are no particular restrictions on the length of an answer-string (which is normally very short), but unnecessary pieces of information will be penalized, since the answer will be marked as non-exact. The answer string must contain nothing more than a complete and exact answer, i.e. *the minimum amount of information needed to satisfy the query.*

Only one exact answer is allowed. Each exact answer string must be accompanied by the following information:

1. The document [*doc_id*] of the ACQUIS or EUROPARL corpora in which it is contained;
2. The id of the paragraph [*p_id*] from which the answer has been cut/pasted;
3. The text of the paragraph from which the exact answer has been extracted.

V. I THE NOA ANSWER

Just as in the 2009 ResPubliQA campaign, systems once again have the option of withholding their answer to a question because they are not sufficiently confident that it is correct. In such a case they can return NOA (i.e. NO Answer). This is not to be confused with the traditional NIL response which in previous campaigns was returned by a system which was unable to find an answer.

The idea behind NOA is that systems can improve their performance by maintaining the number of correct answers in a run while at the same time reducing the number of incorrect answers. The decision as to whether an answer should be returned or whether NOA should be returned is normally taken by an Answer Validation component. In

many cases these components are implemented using machine learning algorithms.

The evaluation measure used this year (see later) is designed to reward a system which returns NOA in those cases where the returned response was in fact wrong. However, a system choosing to leave some questions unanswered must ensure that the number of right answers is maintained, otherwise the resulting score could be lower than it would have been if NOA had not been used in the run.

In summary, the evaluation measure rewards the correct use of NOA and punishes the incorrect use of NOA.
In addition to the above, systems that return NOA for a particular question can also optionally include an answer. In this case, the run will be given two scores, one using the NOA responses and one ignoring NOA and using instead the answers returned. In this way, it will be possible to measure the performance of a system's Answer Validation component.

VI. FORMATS

VI. I TEST SET FORMAT

The DTD for the test sets format can be downloaded from the ResPubliQA web site http://celct.isti.cnr.it/ResPubliQA/Downloads

Test sets will be formatted as an xml file (UTF-8 encoded). The xml will be structured with elements containing the following information:

source_lang target_lang q_id q_string

where:

- **source_lang** is the source language
- **target_lang** is the target language
- **q_id** is the question number (4 digits – 0001 to 0200)
- **q_string** is the question (UTF-8 encoded) string

i.e.:

```
<?xml version="1.0" encoding="UTF-8" ?>
<input>
      <q q_id="0001-0200"
      source_lang="BG|DE|EN|ES|FR|IT|NL|PT|RO"
      target_lang="BG|DE|EN|ES|FR|IT|NL|PT|RO">Question?</q>
</input>
```

Example:

Four questions in a hypothetical EN-EN test set – i.e. English questions that hit the English document collection - might be represented as follows:

```
<?xml version="1.0" encoding="UTF-8" ?>

<input>

      <q q_id="0001" source_lang="EN" target_lang="EN"> What should the
      driver of a Croatian heavy goods vehicle carry?</q>
      <q q_id="0002" source_lang="EN" target_lang="EN"> What should the
      Commission under Regulation (EC) No 2422/2001 create?</q>
      <q q_id="0003" source_lang="EN" target_lang="EN"> What convention
      was done at Brussels on 15 December 1950?</q>
      <q q_id="0004" source_lang="EN" target_lang="EN"> What is another
      name for rights of transit?</q>

</input>
```

VI. II SUBMISSION FORMAT

The DTD for the submission format can be downloaded from the ResPubliQA web site http://celct.isti.cnr.it/ResPubliQA/Downloads

A submission file for the **PS task** must be an xml file (**UTF-8 encoded)** in the form:

```
<?xml version="1.0" encoding="UTF-8" ?>

<output>
      <task_PS>
            <a    q_id="0001-0200"  run_id="XXXX101PSXXXX"
      answered="YES|NO">
            <passage_string      p_id="11"      docid      "jrc31960D051-
      en.xml">xyz</passage_string>
            </a>
      </task_PS>
</output>
```

where:

- **q_id** is the question number as given in the test set (of the form 0001 to 0200) Passages must be returned in the same ascending (increasing) order in which questions appear in the test set
- **run_id** is an alphanumeric string which identifies the runs of each participant. It should be the concatenation of the following elements:
  - ~ the **team ID** (sequence of four lower case ASCII characters)
  - ~ the **current year** (10 stands for 2010)
  - ~ the **number of the run** (1 for the first one, or 2 for the second one)
  - ~ the **task identifier** (PS or AS), obviously, once the task has been selected each question must receive the same kind of response
  - ~ the **language pairs** including both source and target languages, as in the test set.

  Clearly, the content of this field never changes within the same submission file. Each submission file must be named, with an .xml extension, e.g. "clct101PSitit.xml"
- **answered** indicates if question has been answered or not.
- **passage_string** the entire paragraph of text which encloses the answer to the question
- **p_id** is the number of the paragraph from which the paragraph has been extracted
- **docid** is the ID of the document

A sample submission file is reproduced here below, which consists of the passage strings found by human assessors.

Example:

```
- <output>

<task_PS>
- <a q_id="0001" run_id="clct101PSenen" answered="YES">
  <passage_string p_id="21" docid="jrc22003A0618_01-en.xml">4. The driver of a Croatian heavy goods vehicle registered on or after 1 October 1990 shall also carry, and produce upon request, a COP document, modelled on Annex E, as evidence of the NOx emissions of that vehicle. Heavy goods vehicles first registered before 1 October 1990 or in respect of which no document is produced shall be assumed to have a COP value of 15,8 g/kWh.</passage_string>
  </a>
- <a q_id="0002" run_id="clct101PSenen" answered="YES">
  <passage_string p_id="10" docid="jrc32003D0168-en.xml">(1) In compliance with Regulation (EC) No 2422/2001, the Commission should establish a European Community Energy Star Board (hereinafter referred to as the "ECESB" to carry out the EC Energy Star programme, as defined in the Agreement between the Government of the United States of America and the European Community on the coordination of energy efficient labelling programmes for office equipment(2).</passage_string>
  </a>
- <a q_id="0003" run_id="clct101PSenen" answered="YES">
```

```
  <passage_string p_id="8" docid="jrc21987A0720_01-en.xml">CONSIDERING that changes
    in technology and the patterns of international trade require extensive modifications to
    the Convention on Nomenclature for the Classification of Goods in Customs Tariffs, done
    at Brussels on 15 December 1950,</passage_string>
   </a>
- <a q_id="0004" run_id="clct101PSenen" answered="YES">
   <passage_string p_id="7" docid="jrc22003A0618_01-en.xml">1. Ecopoints (rights of
    transit) for Croatian heavy goods vehicles transiting through Austria allocated for 2003:
    171904 ecopoints.</passage_string>
   </a>
   </task_PS>

  </output>
```

A submission file for the **AS task** must be an xml file (**UTF-8 encoded)** in the form:

```
<?xml version="1.0" encoding="UTF-8" ?>

<output>
      <task_AS>
            <a    q_id="0001-0200"  run_id="XXXX101ASXXXX"
      answered="YES|NO">
                  <passage_string     p_id="11"     docid     "jrc31960D051-
            en.xml">xyz
                  </passage_string>
                  <exact_answer> xyz </exact_answer>
            </a>
      </task_AS>
</output>
```

where:

- **q_id** is the question number as given in the test set (of the form 0001 to 0200) Passages must be returned in the same ascending (increasing) order in which questions appear in the test set
- **run_id** is an alphanumeric string which identifies the runs of each participant. It should be the concatenation of the following elements:
  - ~ the **team ID** (sequence of four lower case ASCII characters)
  - ~ the **current year** (10 stands for 2010)
  - ~ the **number of the run** (1 for the first one, or 2 for the second one)
  - ~ the **task identifier** (PS or AS), obviously, once the task has been selected each question must receive the same kind of response
  - ~ the **language pairs** including both source and target languages, as in the test set.

  Clearly, the content of this field never changes within the same submission file. Each submission file must be named, with an .xml extension, e.g. "clct101PSitit.xml"
- **answered** indicates if question has been answered or not

- **passage_string** is the entire paragraph of text which encloses the answer to the question
- **p_id** is the number of the paragraph from which the passage string has been extracted
- **docid** is the ID of the document
- **exact_answer** is the shortest string of text which contains a complete answer to the question

A sample submission file is reproduced here below, which consists of the passage strings found by human assessors.

Example:

```
- <output>

<task_AS>
- <a q_id="0001" run_id="clct101ASenen" answered="YES">
  <passage_string p_id="21" docid="jrc22003A0618_01-en.xml">4. The driver of a Croatian heavy goods vehicle registered on or after 1 October 1990 shall also carry, and produce upon request, a COP document, modelled on Annex E, as evidence of the NOx emissions of that vehicle. Heavy goods vehicles first registered before 1 October 1990 or in respect of which no document is produced shall be assumed to have a COP value of 15,8 g/kWh.</passage_string>
    <exact_answer>a COP document</exact_answer>
  </a>
- <a q_id="0002" run_id="clct101PSenen" answered="YES">
   <passage_string p_id="10" docid="jrc32003D0168-en.xml">(1) In compliance with Regulation (EC) No 2422/2001, the Commission should establish a European Community Energy Star Board (hereinafter referred to as the "ECESB" to carry out the EC Energy Star programme, as defined in the Agreement between the Government of the United States of America and the European Community on the coordination of energy efficient labelling programmes for office equipment(2).</passage_string>
  <exact_answer>a European Community Energy Star Board</exact_answer>
  </a>
- <a q_id="0003" run_id="clct101PSenen" answered="YES">
  <passage_string p_id="8" docid="jrc21987A0720_01-en.xml">CONSIDERING that changes in technology and the patterns of international trade require extensive modifications to the Convention on Nomenclature for the Classification of Goods in Customs Tariffs, done at Brussels on 15 December 1950,</passage_string>
  <exact_answer>the Convention on Nomenclature for the Classification of Goods in Customs Tariffs</exact_answer>
  </a>
- <a q_id="0004" run_id="clct101PSenen" answered="YES">
   <passage_string p_id="7" docid="jrc22003A0618_01-en.xml">1. Ecopoints (rights of transit) for Croatian heavy goods vehicles transiting through Austria allocated for 2003: 171904 ecopoints.</passage_string>
  <exact_answer>Ecopoints </exact_answer>
  </a>
    </task_AS>

  </output>
```

VII. EVALUATION

Systems are allowed to participate in one or both tasks (PS and/or AS) which will operate simultaneously on the same input questions. A maximum of two runs in total can be submitted, i.e. two PS runs, two AS runs or one PS plus one AS run.

Only one response per question will be permitted this year.

Each run for both the PS and AS tasks will be **automatically** evaluated against the Gold Standard manually produced.

Non-matching paragraphs and answers will be manually evaluated by native speaker assessors.

PARAGRAPH SELECTION TASK

Human assessors will have to decide among three kind of judgements:

- **R** (Right): The paragraph returned contains a correct answer;

- **W** (Wrong): The paragraph returned does not contain a correct answer;

- **U** (Unanswered): The system gave NOA.

ANSWER SELECTION TASK

Human assessors will consider correctness (i.e. responsiveness) and exactness (i.e. the quantity of information) of the returned answers.

Each non-matching paragraph of the submitted runs will be assessed and marked with one of the following judgments:

- **R** (Right): The answer-string consists of an exact and correct answer, supported by the returned paragraph;

- **X** (ineXact): The answer-string contains either part of a correct answer present in the returned paragraph or it contains all the correct answer plus unnecessary additional text;

- **M** (Missed): The answer-string does not contain a correct answer even in part but the returned paragraph in fact does contain a correct answer. In other words,

the answer was there but the system missed it completely (i.e. did not extract it correctly);

- **W** (Wrong): The answer-string does not contain a correct answer and moreover the returned paragraph does not contain it either;

- **U** (Unanswered): No answer is provided by the system.

The main measure considered in this evaluation campaign will be the following:

$$c@1 = \frac{1}{n}(n_R + n_U \frac{n_R}{n})$$

where

n_R: is the number of correctly answered questions
n_U: number of unanswered questions
n: the total number of questions

Notice that this measure is parallel to the traditional accuracy used in past editions. The interpretation of the measure is the following:

1. a system that gives an answer to all the questions receives a score equal to the accuracy measure used in the previous QA@CLEF main task: in fact, since in this case $n_U = 0$ then $c@1 = n_R/n$;

2. the unanswered questions add value to c@1 only if they do not reduce much the accuracy (i.e. n_R/n) that the system would achieve responding to all questions. This can be thought of as a hypothetical second chance in which the system would be able to replace some NoA answers by the corrects one;

3. a system that does not respond to any question (i.e. returns only NOA as answer) receives a score equal to 0, as $n_R=0$ in both addends.

The adoption of the c@1 evaluation metric encourages systems to maintain the number of correct answers while reducing the amount of incorrect ones by leaving some questions unanswered (NOA). Answer Validation techniques (including Machine Learning) are expected to be used for taking this final decision.

Regarding the evaluation of **exact answers** and their supporting paragraphs, a measure of the performance achieved in extracting the answers once the paragraph has been selected will be provided. That is, both paragraph and exact answer will be assessed. For the evaluation of exact answers, five judgements will be used, as already discussed: Right, ineXact, Missed, Wrong and Unanswered. In addition to computing the c@1 score, we can also measure how well the system extracts answers from paragraphs:

answer extraction performance = #R / (#R + #X + #M)

ORGANIZING COMMITTEE

Anselmo Peñas (UNED, Spain) Co-chair

Pamela Forner (CELCT, Italy) Co-chair

Richard Sutcliffe (University of Limerick, Ireland) Co-chair

Corina Forascu (UAIC and RACAI, Romania)

Álvaro Rodrigo (UNED, Spain)

ADVISORY BOARD

Donna Harman (National Institute for Standards and Technology (*NIST*), USA)

Maarten de Rijke (University of Amsterdam, The Netherlands)

Dominique Laurent (*Synapse* Dévelopement, France)

CONTACT INFORMATION

Anselmo Peñas: anselmo *at* lsi.uned.es

Pamela Forner: forner *at* celct.it

Richard Sutcliffe: richard.sutcliffe *at* ul.ie