# CLEF Guidelines - AdHoc 2004

## Test Collection

The main document collection in CLEF 2004 has comparable documents for national newspapers and newswires in 10 languages: Dutch, English, Finnish, French, German, Italian, Portuguese, Russian, Spanish, Swedish. It has been enlarged with respect to the CLEF 2003 collection with a new language (Portuguese) and additional data.

The time span now covered for most languages is 1994-95, with the exception of the Finnish (Nov. 94-through 95) and Russian (1995 only) collections. The mono-, bi- and multilingual tracks use only the English, Finnish, French, Portuguese and Russian collections for 1995.

The topic set for the Ad Hoc tasks consists of 50 topics and is prepared in: Dutch, English, Finnish, French, German, Italian, Portuguese, Russian, Spanish, Swedish (main languages) and Japanese (additional topic language). The original topic set is prepared in EN, FI, FR, PT, RU on the basis of the contents of the collections and consists of a selection of topics of local (i.e. national), European, and general interest. This means that the number of relevant documents in any one collection can vary considerably, depending on the topic; in some cases, for a particular topic, there may be no relevant documents in a given collection. Guidelines in other European languages are translated from the original topics by independent translators (i.e. not belonging to participating groups). Japanese topics were also prepared to encourage the participation of Asian groups interested in European languages and vice versa.

Guidelines are released using the correct diacritics (according to the language) but may contain occasional spelling errors/inconsistencies, minor formatting deficiencies. We aim to keep these at a minimum. Since the documents for the CLEF experiments come from well-known high-quality sources, they should have a very small error-rate with respect to accents. However, participants should still be prepared for accent mismatches. This constitutes a real-world problem. Note that accents may be transcribed if this is common practice in the area of origin of the documents. In particular, the accents in one of the Italian collections (La Stampa) are indicated by a following apostrophe (') as a final character whereas in the other (SDA) the correct accented chracters are used. The English documents are from both the US and the UK. Depending on the track in which you participate and the collection(s) used there can thus be lexical and orthographic mismatches. The German collection also contains documents in German using Swiss-specific vocabulary. Systems must be sufficiently robust to cater for such features (very common in real world situations).

## Tasks

The goals of the tasks are as follows:

- **Multilingual**: Selecting a topic set in any language, retrieve relevant documents from the multilingual collection of English, Finnish, French, and Russian news documents for 1995 and submit the results in a single ranked list.

- **Bilingual**: The 2004 bilingual track on news collections will accept runs for the following source -> target language pairs:
  - Italian/French/Spanish/Russian topics -> Finnish target collection
  - German/Dutch/Finnish/Swedish topics -> French target collection
  - Any topic language -> Russian target collection
  - Any topic language -> Portuguese target collection

  Newcomers only (i.e. groups that have not previously participated in a CLEF cross-language task) can choose to search the English document collection using any topic language. The aim is to retrieve relevant documents from the chosen target collection and submit the results in a ranked list.
- **Monolingual (non-English)**: Query the Finnish, French, Portuguese or Russian collections using topics in the same language and submit results in a ranked list.

Much of the evaluation methodology is an adaptation of the strategy studied for the TREC ad-hoc task. The instructions given below have been derived from those distributed by TREC.

# CONSTRUCTING AND MANIPULATING THE SYSTEM DATA STRUCTURES

The system data structures are defined to consist of the original documents, any new structures built automatically from the documents (such as inverted files, thesauri, conceptual networks, etc.), and any new structures built manually from the documents (such as thesauri, synonym lists, knowledge bases, rules, etc.).

1. The system data structures may not be modified in response to CLEF 2004 topics. For example, you cannot add topic words that are not in your dictionary. The CLEF tasks represent the real-world problem of an ordinary user posing a question to a system. In the case of the cross-language tasks, the question is posed in one language and relevant documents must be retrieved whatever the language in which they have been written. If an ordinary user could not make the change to the system, you should not make it after receiving the topics.

2. There are several parts of the CLEF data collections that contain manually-assigned, controlled or uncontrolled index terms. These fields are delimited by SGML tags. Since the primary focus of CLEF is on retrieval of naturally occurring text over language boundaries, these manually-indexed terms should not be indiscriminately used as if they are a normal part of the text. If your group decides to use these terms, they should be part of a specific experiment that utilizes manual indexing terms, and these runs should be declared as manual runs.

3. Only the following fields may be used for automatic retrieval (collections used in CLEF2004 adhoc tracks are in red):

- Frankfurter Rundschau: TEXT, TITLE only
- Der Spiegel: TEXT, LEAD, TITLE only
- La Stampa: TEXT, TITLE only
- Le Monde 95: TEXT, LEAD1, TITLE only
- LA TIMES: HEADLINE, TEXT only
- Glasgow Herald: HEADLINE, TEXT only

- NRC Handelsblad: P only (or alternatively TI, LE, TE, OS only)
- Algemeen Dagblad: P only (or alternatively TI, LE, TE, OS only)
- EFE 94 and 95: TITLE, TEXT only
- German/French/Italian SDA 94 and 95: TX, LD, TI, ST only
- Aamulehti: TEXT only
- TT - Tidningarnas Telegrambyrå: BRODTEXT, HEADLINE, MELLIS, INGRESS, RUBRIK (And of course the tags contained therein. Most of these contain one or more <P> tags. Note however, that <P> tags can also occur elsewhere in the document. You cannot blindly index all <P> tags.)
- Izvestia: TEXT, TITLE only
- Público 95: TEXT only

Learning from (e.g. building translation sources from) such fields is permissible.

# GUIDELINES FOR CONSTRUCTING THE QUERIES

The queries are constructed from the topics. Each topic consists of three fields: a brief title statement; a one-sentence description; a more complex narrative specifying the relevance assessment criteria. Queries can consist of 1 or more of these fields.

There are many possible methods for converting the supplied topics into queries that your system can execute. We have broadly defined two generic methods, "automatic" and "manual", based on whether manual intervention is used or not. When more than one set of results are submitted, the different sets may correspond to different query construction methods, or if desired, can be variants within the same method.

The manual query construction method includes BOTH runs in which the queries are constructed manually and then run without looking at the results AND runs in which the results are used to alter the queries using some manual operation. The distinction is being made here between runs in which there is no human involvement (automatic query construction) and runs in which there is some type of human involvement (manual query construction). It is clear that manual runs should be appropriately motivated in a CLIR context, e.g. a run where a proficient human simply translates the topic into the document language(s) is not what most people think of as cross-language retrieval.

To further clarify this, here are some example query construction methodologies, and their correct query construction classification. Note that these are only examples; many other methods may be used for automatic or manual query construction.

1. queries constructed automatically from the topics, the retrieval results of these queries sent to the CLEF results server --> automatic query construction

2. queries constructed automatically from the topics, then expanded by a method that takes terms automatically from the top 30 documents (no human involved) --> automatic query construction

3. queries constructed manually from the topics, results of these queries sent to the CLEF results server --> manual query construction

4. queries constructed automatically from the topics, then modified by human selection of terms suggested from the top 30 documents --> manual query construction

Note that by including all types of human-involved runs in the manual query construction method we make it harder to do comparisons of work within this query construction method. Therefore groups are strongly encouraged to determine what constitutes a base run for their experiments and to do these runs (officially or unofficially) to allow useful interpretations of the results. For those of you who are new to CLEF, unofficial runs are those not turned into CLEF but evaluated using the trec_eval package available from Cornell University. (See previous years' CLEF papers for examples of use of base runs.)

# WHAT TO DO WITH YOUR RESULTS

Results have to be formated in ASCII, with one line per document retrieved. The lines have to be formatted as follows:

| Field#1 | Field#2 | Field#3 | Field#4 | Field#5 | Field#6 |
|---|---|---|---|---|---|
| 10 | Q0 | document.00072 | 0 | 0.017416 | runidex1 |

The fields must be separated by ONE blank and have the following meanings:

1. Query number (eliminate any identifying letters). Please only use SIMPLE numbers ("1", not "001")
   **INPUT MUST BE SORTED NUMERICALLY BY QUERY NUMBER**.
2. Query iteration (will be ignored. Please choose "Q0" for all experiments).
3. Document number (content of the tag.).
4. Rank 0..n (0 is best matching document. If you retrieve 1000 documents per query, rank will be 0..999, with 0 best and 999 worst). Note that rank starts at 0 (zero) and not 1 (one).
   **MUST BE SORTED IN INCREASING ORDER PER QUERY**.
5. RSV value (system specific value that expresses how relevant your system deems a document to be. This is a floating point value. High relevance should be expressed with a high value). If a document D1 is considered more relevant than a document D2, this must be reflected in the fact that RSV1 > RSV2. If RSV1 = RSV2, the documents may be randomly reordered during calculation of the evaluation measures. Please use a decimal point ".", not a comma. Do not use any form of separators for thousands. The only legal characters for the RSV values are 0-9 and the decimal point.
   **MUST BE SORTED IN DECREASING ORDER PER QUERY**.
6. Run identifier (please chose an unique ID for each experiment you submit). Only use a-z, A-Z and 0-9. No special characters, accents, etc. The fields are separated by a single space. The file contains nothing but lines formatted in the way described above. You are expected to retrieve 1000

documents per query. An experiment that retrieves a maximum of 1000 documents each for 20 queries therefore produces a file that contains a maximum of 20000 lines.

You should know that the effectiveness measures used in CLEF evaluate the performance of systems at various points of recall. Participants must thus return at most 1000 documents per query in their results. Please note that by its nature, the average precision measure does not penalize systems that return extra irrelevant documents at the bottom of their result lists.