Project no. 215231

TrebleCLEF

Treble-CLEF: Evaluation, Best Practices and Collaboration for Multilingual Information Access
IST: ICT-1-4-1, Digital libraries and technology-enhanced learning

Deliverable 3.3
Best Practices in System-oriented and User-oriented Multilingual Information Access

Start Date of Project: 01 January 2008
Duration: 24 Months

Organisation Name of Lead Contractor for this Deliverable: UNED

Version 2.0, September 15, 2009

Project co-funded by the European Commission within the Seventh Framework programme

Document Information

Deliverable number:	3.3
Deliverable title:	Best Practices in System-oriented and User-oriented Multilingual Information Access
Due date of deliverable:	18/09/2009
Actual date of deliverable:	18/09/2009
Author(s):	Martin Braschler & Julio Gonzalo
Participant(s):	ZHAW, UNED
Workpackage:	3
Workpackage title:	Best Practices in System Development and User Studies
Workpackage leader:	UNED
Dissemination Level:	Public
Version:	2.0
Keywords:	best practices, information retrieval, cross-language information retrieval, multilingual information access, system-oriented

History of Versions

Version	Date	Status	Author (Partner)	Description/Approval Level
1.0-	25/06/2009	Preliminary & partial	ZHAW, UNED	Description of system-oriented best practices and partial version of user-oriented aspects
2.0	24/09/2009	Final	Martin Braschler (ZHAW), Julio Gonzalo, Víctor Peinado, Fernando López-Ostenero (UNED)	Full deliverable

Abstract

While there is growing interest in multilingual information access (MLIA) in a range of fields, only few operational systems exist. MLIA systems typically incorporate technologies and tools from a number of areas: information retrieval, natural language processing, language resources and potentially others. They thus remain complex to implement. In the past decade, there has been increasing uptake of MLIA and CLIR (cross-language information retrieval) problems in the academic community, and the result is an expanding body of research literature, much of it produced in the confines of CLEF (Cross-Language Evaluation Forum). We intend to facilitate the transfer of MLIA/CLIR technology by highlighting "unifying" conclusions from these academic experiments, and compiling them into best practice recommendations. The recommendations are based on the analysis of papers from CLEF and related venues, as well as the consideration of the results from a pair of workshops attended by experts in operational MLIA. This deliverable summarizes the recommendations for the areas of system-oriented and user-oriented multilingual information access. We show that there are indeed basic building blocks that can be used to obtain robust MLIA systems in many situations, as demonstrated through evaluation at CLEF. A full table of the recommendations is given at the end of the deliverable.

Table of Contents

Executive Summary

This deliverable presents best practice recommendations for system-oriented and user-oriented multilingual information access (MLIA). We intend to facilitate the transfer of MLIA/CLIR (cross-language information retrieval) technology by highlighting "unifying" conclusions from academic experiments (mainly conducted within the CLEF (Cross-Language Evaluation Forum) campaigns and related venues). Additionally, the results from a pair of workshops attended by experts in the field of operational MLIA were also consulted.

Main recommendations for system-oriented MLIA are:

1. Use retrieval systems supporting term weighting and ranked retrieval.
2. Use minimal stopword (non-content bearing word) elimination; remove special characters (diacritics).
3. Use stemming (word form reduction), use decompounding for languages with productive compound formation.
4. If 3 is impossible, due to missing resources for certain languages, use character n-grams.
5. Maximize the coverage of translation resources, potentially by combining multiple resources, and add domain-specific resources
6. Combine different types of translation resources, if computational and financial costs are acceptable
7. Use one of a set of high-performing, well-researched weighting schemes for ranking results
8. Use pseudo-feedback as a query enhancement technique if recall is of concern

Main recommendations for user-oriented components of MLIA systems are:

1. To facilitate document selection, use cross-language summaries or high-quality MT.
2. If feasible, translate the whole document collection at index time for simpler query formulation and reformulation issues.
3. Include user-assisted query translation facilities, but do not show them by default.
4. Indirect user-assisted query translation that does not involve inspecting foreign-language terms is preferable.
5. Design document translation and query translation/refinement facilities to fit together.
6. Combine text-based with content-based facilities for cross-language image search.
7. If feasible, use monolingual IR over a translated document collection as backbone for CL-QA systems.

See the summary of the deliverable for a full list of all recommendations, including justifications.

1 Introduction

Despite growing interest in multilingual information access (MLIA) in a range of fields, such as digital libraries, enterprise search and web applications, there still exist few operational systems. One of the core challenges of multilingual information access is the problem of cross-language information retrieval (CLIR): how to access documents written in any one (or even more than one) of a range of different languages, given a query formulated in the language of the user's preference; or, from a broader, user-inclusive perspective, how best to assist users finding information regardless of the mismatches between the user's language skills and the languages in the document collection. The MLIA system must bridge the language gap between the documents and the user's request in an appropriate manner. Clearly, faced with today's ever-growing digital universe, the ability to effectively and efficiently access information in many languages can be a crucial competitive advantage. To illustrate this point, IT market research firm IDC forecasts growth in 2009 for multilingual/cross-language applications and tools, likely fueled by the need to move into emerging economies[1] .

An MLIA system thus has to mediate between different languages, normally by employing some form of translation. The lack of commercial uptake of MLIA technology is in contrast to a large and ever expanding body of academic work in the fields of MLIA and CLIR. Much of this work has been conducted by participants in the CLEF[2] (cross-language evaluation forum) evaluation campaigns. To date, there have been 9 yearly CLEF campaigns, starting with the first campaign in 2000. A total of around 200 different groups have participated in these campaigns, and have compiled several hundred different papers describing their evaluation experiments. Roughly 90% of participants can be classified as "academic institutions" (mainly universities).

The implementation of a fully operational multilingual or cross-language information retrieval system remains complex and involves integrating technologies and tools from a number of areas: Information retrieval, Natural Language Processing, Language Resources, Computer-Human Interaction. Currently, there are no easily applicable off-the-shelf solutions for new players that want to move into the field. While the evaluation focus of CLEF ensures that the experiments are highly interesting for practitioners that want to understand the state-of-the-art, the resulting body of literature can be daunting to access. As part of the efforts of the TrebleCLEF coordination action[3], we want to facilitate the transfer and uptake of innovative MLIA/CLIR technology. We cannot, by the nature of our efforts as part of TrebleCLEF, package MLIA components ourselves. However, we hope that the analysis of academic experiments contained in this report can direct interested practitioners to the relevant approaches.

Objectives

The primary objective is to provide recommendations for successful CLIR/MLIA implementations, generalizing as much as possible from the mostly academic papers.

There are few "packaged" MLIA/CLIR offerings in the marketplace, and the guidelines alone cannot close this gap. Instead, we hope the recommendations are found to be valuable for commercial developers/implementers of complex search applications, which are ready to build their own components that are tailored to those specific applications.

It is outside of the scope of the work of the TrebleCLEF coordination action to conduct new experiments, or to develop tools and components. Instead, the findings in this paper harness the existing experiment descriptions submitted by the individual participants in CLEF.

[1] Gens, F.: IDC Predictions 2009
[2] http://www.clef-campaign.org
[3] http://www.trebleclef.eu

The main challenge is to find a "unification" of the conclusions from a vast range of different experiments, many of which use different (not always explicitly stated) testing hypotheses, one of the varied CLEF test collections (data and queries) and methodologies (ad-hoc vs. interactive tasks) and different system parameters. The results from the CLEF campaigns have shown that there are multiple competing approaches, which can solve the MLIA/CLIR problem to a differing degree given the varying experiments. While we hope that practitioners will find the recommendations helpful, the report cannot replace the analysis of further literature (some of which is listed in the references), and there may well be alternative approaches. It is inevitable that there may not be unanimous agreement on all the recommendations that we give. We hope to help further analysis by clearly indicating what the recommendation is based on in each case. The report is slated to be published on the TrebleCLEF portal at a later date, and we hope that both the academic community and practitioners will provide much appreciated commentary, corrections and extensions to this text.

The report is organized in two broad aspects of a MLIA system: (i) the ad-hoc retrieval component, which has queries as input and documents as output; and (ii) the interactive aspects of an MLIA process, i.e. how best to assist users formulating, translating and reformulating queries, how best to assist users detecting document relevance in foreign languages, how to facilitate optimal user feedback, etc. The first of these two aspects of an MLIA system, covered in Section 2, is the one that has received, by far, most attention in CLEF, and this report summarizes hundreds of contributions in this area plus input from a TrebleCLEF workshop on operational MLIA. The second aspect, i.e. user-oriented aspects of MLIA, has only been the focus of a small CLEF track (iCLEF), but its findings are probably crucial for developers interested in implementing fully functional, effective multilingual search assistants. In Section 3 we report findings from iCLEF and complement them with the outcome of another TrebleCLEF workshop which gathered best-practice recommendations from relevant user communities.

2 System-oriented MLIA

In this section, we present an analysis of the main components that implement the MLIA/CLIR flow in the "ad-hoc" retrieval sense, i.e. they are used by the systems to process the query (input) and return a set of matching documents (output). We will discuss the role of these components as a part of a large information acquisition cycle.

2.1 Best Practices in System-oriented MLIA

2.1.1 General requirements

Many of the recommendations contained in this report are heavily based on the results reported by the participants in the CLEF[4], TREC[5] and NTCIR[6] evaluation campaigns. We mostly concentrate on the CLEF experiments, where the bulk of research in cross-language information retrieval for European languages is reported. We extend this coverage to TREC and NTCIR where appropriate.

The experiments conducted in the field of multilingual information access and cross-language information retrieval at the CLEF campaign use nearly exclusively systems that present search results in the form of ranked results lists sorted by descending order of probability of relevance. Queries, the formulations of information need by the users, are given in natural-language form (either as a set of keywords or as well-formed, grammatical sentences), while matches between queries and documents are based on full-text search using term weighting. This form of retrieval system is a good fit for the inherently ambiguous nature of translation between languages. There are very few exceptions in CLEF in the form of experiments that use different retrieval paradigms such as Boolean retrieval - see e.g. (Ripplinger 2000).

Condensing the findings of the successful CLEF experiments, the report will consequently concentrate on such approaches for ranked full-text retrieval, representing this broad consensus of the CLEF academic community. The focus on text retrieval derives from the especially large corpus of academic work dealing in MLIA/CLIR with "ad-hoc" text retrieval. The "ad-hoc" track at CLEF has always been considered the "core track" (see e.g. (Agirre et al. 2008)). Related information access problems such as question answering and image retrieval would easily deserve a full best practices report of their own and the specifics of these specialized fields are therefore outside the scope of this report. Note, however, that cross-language "ad-hoc" text retrieval (or "document retrieval") is the major backbone for most information access tasks, and therefore the recommendations are, in general, applicable to most MLIA tasks.

Recommendation	Based on
Use a retrieval system supporting term weighting and ranked retrieval. This form of retrieval system addresses the inherently ambiguous nature of translation between languages well	Necessary pre-condition for most of the state-of-the-art CLIR/MLIA components

Information Retrieval systems such as those outlined above make use of many components that address problems also analyzed in the academic field of computational linguistics. Specifically, words need to be extracted from queries and documents, and need to be normalized for effective matching. Both queries and documents can be analyzed with linguistic methods to improve matching. Please note, however, that the primary goal of information retrieval is an optimization of the retrieval process, i.e. the effectiveness and efficiency of retrieval, and not of linguistic processing. The two viewpoints,

[4] http://www.clef-campaign.org
[5] http://trec.nist.gov
[6] http://research.nii.ac.jp/ntcir/

optimized retrieval and linguistic correctness, can very well be in conflict, with linguistic analysis adding "noise" to the retrieval process, as we will note in the corresponding discussions of components that are affected.

Information Retrieval systems are part of a larger "information acquisition" cycle (see Figure 1). It is fair to assume that interaction with the system is only part of a larger process initiated by the user to acquire some information necessary to solve a problem or satisfy an information need. Please note the fundamental difference to database systems and retrieval from databases: in databases, correctly structured data is stored according to a well-defined database schema. In information retrieval, users try to satisfy information needs by querying an unstructured database, potentially fed with information from very disparate sources. Typically, the understanding of the problem or the information need is very incomplete at the outset. The process of working with the information retrieval system becomes an iterative one, as users gain a better insight into possible solutions during querying. In the following sections on system-oriented best practices, we will mainly concentrate on the information retrieval system in the narrower sense, i.e. on the system that takes a coded input query and returns a list of documents. The sections on best practices in user-oriented MLIA address many of the other main points of the information acquisition cycle.

Let us consider a small example in order to better identify the role of the information retrieval system. Suppose the user wants to know how to best archive digital photographs. We assume that she has this information need precisely because she has no solution for this problem in place yet. This poses a paradox: while many relevant documents will likely present descriptions of tools to achieve the desired archival, the user may initially not know the necessary vocabulary to express the query. It is, however, not uncommon that users have a vague idea of a possible solution. This idea can be put into words: we assume the user "realizes" she wants to look into a "CD burner" as an archival solution. With this newly expressed information need, she starts using the information retrieval system proper. The system will present some form of user interface where the query can be typed in. While most systems built by CLEF participants for their experiments allow the use of full natural language queries, many users are conditioned to input a (short) sequence of keywords, stemming from the use of Web search engines such as Google[7] and Yahoo[8], which by default narrow search results by using an implicit "AND" operator for query terms[9]. The user types "CD burner" and starts the search.

The task of the information retrieval system is now to find information relevant to the *initial* information need (archival solutions for digital photographs), which was much wider than the input based on this short query (CD burner). Possible solutions are formulated in a variety of ways: documents can refer to "CDs", "compact discs", "optical discs", maybe also "DVD". The "CD burner" may be called "CD recorder" or "CD-R drive", among other possibilities. And of course, there may be relevant information in documents that do not match the rather narrow verbalization of the initial information need. For example, documents may talk about "flash" devices as an alternative for data storage. Processing the list of results compiled by the system, the user will gain a better understanding about her problem, and will be able to re-verbalize the need – thus enabling a new cycle through the information acquisition process.

[7] http://www.google.com
[8] http://www.yahoo.com
[9] As a consequence, search results will narrow when additional terms are input. Basically, only documents containing all the terms of the query will be shown (although this rule has been weakened lately by Web search engines to allow exceptions, such as the inclusions of documents that do not contain a term, but are linked to by documents that do)

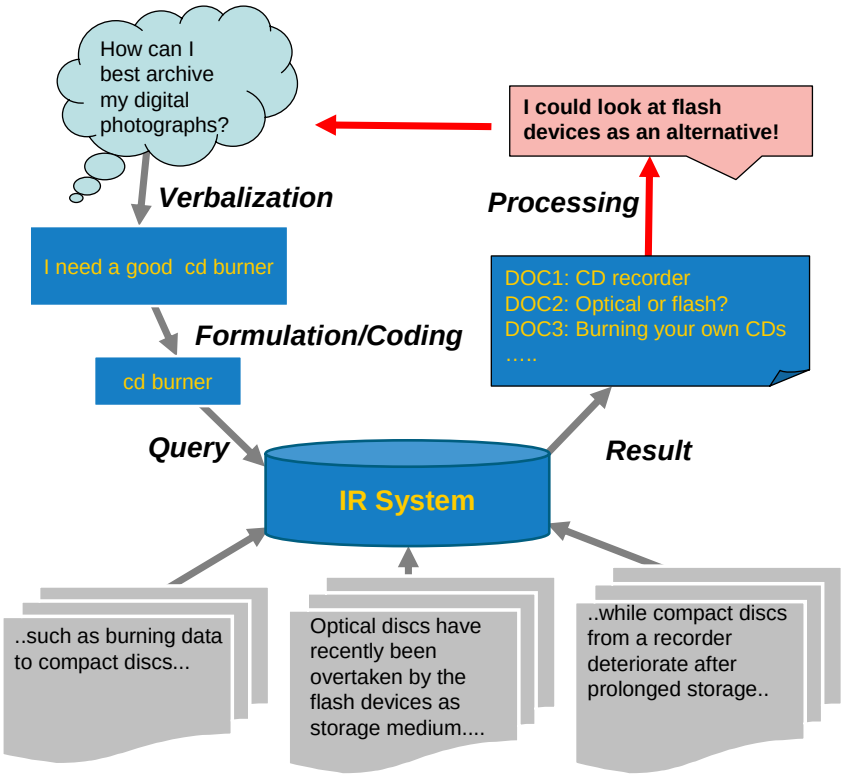

Figure 1: The information retrieval system is part of a larger information acquisition cycle.

2.1.2 Methodology

We base the recommendations given for different components of MLIA/CLIR systems on the results obtained by the various CLEF participants in different CLEF campaigns. In the following, we first provide a short summary of the evaluation methodology used in CLEF.

For the evaluation of textual MLIA/CLIR access to unstructured documents, CLEF has mainly adopted a corpus-based, automated scoring method, based on ideas first introduced in the Cranfield experiments (Cleverdon 1977) in the late 1960s. This allows "lab-style" batch evaluation without direct user involvement. The properties of this method have been thoroughly investigated since its introduction, and are well understood. The same approach is used by all large evaluation campaigns (CLEF, TREC, NTCIR) for many of their activities. For a more detailed discussion of the Cranfield paradigm and its underlying assumptions, see also (Voorhees 2002).

In the CLEF campaigns, a combination of a set of retrievable documents and a set of formulations of "information needs" is used. Two measures are used to evaluate the effectiveness for answering these information needs on the basis of the documents: recall and precision. These two measures model the assumption that users wants to retrieve as much relevant information as possible (recall), while minimizing the amount of irrelevant information also returned (precision) (Schäuble 1997). To compute them, relevance assessments, i.e. human judgments on the relevance of a document with respect to an information need, are necessary. They are also supplied by the CLEF organizers. Documents, information needs, and relevance assessments form a "test collection".

During the CLEF campaign, academic (mainly) and industrial participants benchmark their systems using these test collections while considering a variety of different tasks. It is important to bear in mind that for most of the experiments, average performance over a set of information needs is reported. The use of averages can be problematic, as it hides irregularities on the level of individual queries. We will point out consequences where appropriate in the reminder of this document. The participants report their results in papers that are published in the campaign working notes, as well as in a post-

campaign proceedings volume. These papers contain a wealth of information that is of potential great benefit to practitioners in the MLIA/CLIR field. Participants have a large degree of freedom in how they present their experiments, and thus little structured information about the experiments is available (in databases or elsewhere). It often remains unclear for practitioners how to abstract from individual, concrete experiments in order to generalize to a specific setting currently at hand. Reading through the hundreds of experiment descriptions compiled during nearly a decade of CLEF experiments is a very time-consuming exercise, and it is hard for outsiders to adequately judge the merits of individual experiments.

The results from different CLEF experiments are an important basis for our recommendations. We have attempted to help this process of digesting the CLEF experiment descriptions and generalizing them in an adequate way. To this end, we have not only concentrated on the experiment descriptions, but have used multiple sources of input in order to formulate best practice recommendations that should be applicable to a wider range of MLIA/CLIR scenarios. The sources are:

1. The overview papers of the working notes/proceedings – see e.g. (Agirre et al. 2008), (Braschler 2004a) –. These are summaries of the experiments for specific CLEF tasks conducted within single years. They give a condensed view of some of the results of a single campaign from an academic perspective. Some of them have been (co-)authored by the author of this report over the years, but some have been written by other CLEF organizers.
2. A statistical analysis of the text of the experiment descriptions. We have loaded the raw, unstructured text of the experiment descriptions into an information retrieval system (Lucene[10]). Based on the index built by Lucene, we have extracted lists of characteristic terms using a statistical frequency analysis. These lists give a quick, rough overview over approaches and methods employed in the experiments, and can be used as "seed queries" for interactive searching in the collection of experiment descriptions (also supported by an interactive, cross-referenced collection of CLEF experiment descriptions which we have set up using Lucene). An example is shown in Table 1.
3. The feedback we received by organizing and hosting a workshop on "Best Practices for System Developers: Bringing Multilingual Information Access to Operational Systems". This invitation-only workshop was held in October 2008 in Winterthur, Switzerland and we invited practitioners from the field who had had previous exposure to CLEF and its experiments. Issues when digesting and transferring insights from CLEF experiments were identified, and guidelines derived from practical experience by the participants were formulated at the workshop (Braschler & Clough 2008).
4. An earlier analysis on a successful blueprint for MLIA CLEF experiments we conducted in 2003 (Braschler & Peters 2004b).

In Table 1 we show an example excerpt of the top-ranked terms from one of our term lists which was generated through word frequency analysis (item 2). We have worked through these lists, identifying terms which indicate the use of specific techniques and algorithms by the respective participants. In this sample excerpt, prime candidates are the entries “queri.expans”, “name.entiti” and “relev.feedback”. Please note that the terms are shown in their “stemmed” forms, i.e. after word form reduction (this is also a technique used for MLIA/CLIR, see Section 2.1.6, step 5 below). It is still easy with a good working knowledge of the MLIA/CLIR field to identify the underlying original word forms. We can then use these terms as “seed queries” to interactively explore the collection of experiment descriptions which we have loaded into the Lucene search system. Furthermore, we get a good indication of the frequency with which the method or algorithm has been used in CLEF: the fields cf (collection frequency) and df (document frequency) denote the total number of occurrences of the term in all documents and the number of documents with at least one occurrence of the term, respectively.

[10] http://lucene.apache.org

Term	cf	df
averag.precis	1541	294
cross.languag	1314	338
relev.document	1279	296
queri.expans	1267	238
question.answer	1255	175
document.collect	1186	323
name.entiti	1144	171
imag.retriev	904	107
retriev.system	892	334
submit.run	871	284
relev.feedback	812	207
queri.term	783	216

Table 1: Example of a list of terms generated through frequency analysis of the CLEF experiment descriptions (papers). Shown are the top twelve two-word terms, in their "stemmed" form, along with collection frequency and document frequency. As can be seen, the list gives a good overview of the technical terminology used in the underlying papers, and thus serves as a list of potential search terms for interactive exploration of the papers.

We have compiled a number of these lists, for one-word, two-word and three-word terms, sorted by different criteria, to ensure that we identify a maximum number of potential seed queries. We then manually skim these lists for good entries. We are convinced that this technique helps us ensure a better coverage of the source material in our analysis, while leading to comparatively low overhead for automatically compiling the lists. The terms have also been used to produce an internal, cross-referenced version of the collection of CLEF experiment descriptions, suitable to help our analysis.

2.1.3 Different types of Multilingual Information Access

The terms "multilingual information access" and "cross-language information retrieval" have been used in different contexts in the past. To avoid confusion, we list some of the definitions in the following and discuss if and how the techniques described in this report apply to them.

The report concentrates on multilingual information access in the form of multilingual ad-hoc text retrieval, i.e. methods that deliver lists of search results in response to a spontaneous, "ad-hoc" query by a user. Such a query usually denotes a formulation of an information need by the user, and can be part of a larger knowledge acquisition process as outlined in Section 2.1.1. Some MLIA systems provide support for this larger process, and some of the aspects of such systems are discussed in the section on user-oriented best practices. Insofar as the retrieval system itself can directly support this process, we will discuss relevant techniques as well (see e.g. Section 2.1.8).

When concentrating on this form of multilingual information retrieval, four different forms of MLIA are often mentioned:

1. monolingual access to documents written in languages other than English
2. bilingual access to documents written in a language different from the language used for query formulation ("source language" to "target language")
3. multilingual access to monolingual documents written in any of a number of languages ("target languages"), using a query in the language of the user's preference ("source language")
4. multilingual access to multilingual documents, with the language of the query and the documents drawn from a set of different languages (all of them potentially "source" and "target" languages).

The definitions imply an increasing degree of multilinguality, with definition 4 allowing an almost arbitrary use of languages in documents and queries.

Definition 4 also matches most closely the "Grand challenge" formulated by D. Oard and D. Hull at the AAAI Symposium on Cross-Language IR in 1997: "Given a query in any medium and any language, select relevant items from a multilingual multimedia collection which can be in any medium and any language, and present them in the style or order most likely to be useful to the querier, with identical or near identical objects in different media or languages appropriately identified.".[11]

2.1.4 The MLIA/CLIR "flow"

Structurally, the discussion of a retrieval system providing support for multilingual information access can be segmented into three main phases:

1. indexing phase
2. translation phase
3. matching phase

All retrieval systems of the type covered by this report (systems for ranked retrieval on large datasets) use some form of an index which is pre-built and periodically updated (indexing phase). Searching large datasets without such an index is not practicable, as the search would require linear scans at query execution time, with the corresponding run time for such a scan quickly becoming prohibitive. Once an index is available, retrieval systems use the information stored in the index to look up terms from the user's queries and calculate a score for each matching document (matching phase) (Figure 2).

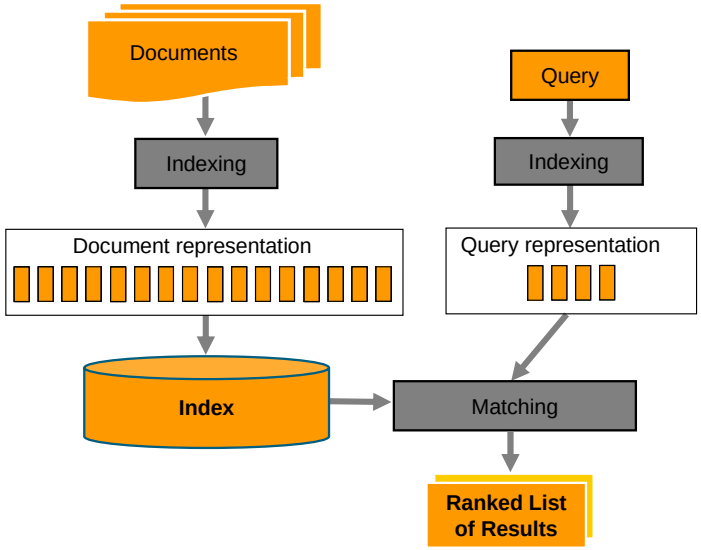

Figure 2: Basic workings of a monolingual information retrieval system. An "indexing" phase converts both documents and queries to an internal representation which is suitable for retrieval in a "matching" phase. The matching phase calculates scores for each document with respect to the query, which are used to produce a ranked list of results.

In the multilingual case, terms from the query will not match terms in the documents if different languages are used. We need an intermediate step designed to bridge this language gap, normally some form of translation (translation phase). Please note, however, that translation in this context is used only in the loose sense of providing a transfer mechanism between languages that is suitable for

[11] See http://terpconnect.umd.edu/~dlrg/filter/sss/ for more information on the symposium

search, and not in the stricter linguistic sense of rendering a text in a new language while preserving the original meaning as accurately as possible.

There is not a single, fixed form of interaction between the three phases. During the translation phase, it is possible to translate either the documents, queries, both, or neither. This choice will influence the combination of the different components for indexing, matching and translation. In the following illustrations (Figure 3, Figure 4, Figure 5), three different possibilities that cover most of the experiments in CLEF are outlined.

Figure 3 shows a bilingual system based on query translation. This is possibly the simplest widely implemented extension of an existing monolingual system. The system translates the query, and subsequently works analogous to the monolingual basic system shown in Figure 2. Note that the "indexing" and "translation" step of the query can also be reversed in some system architectures.

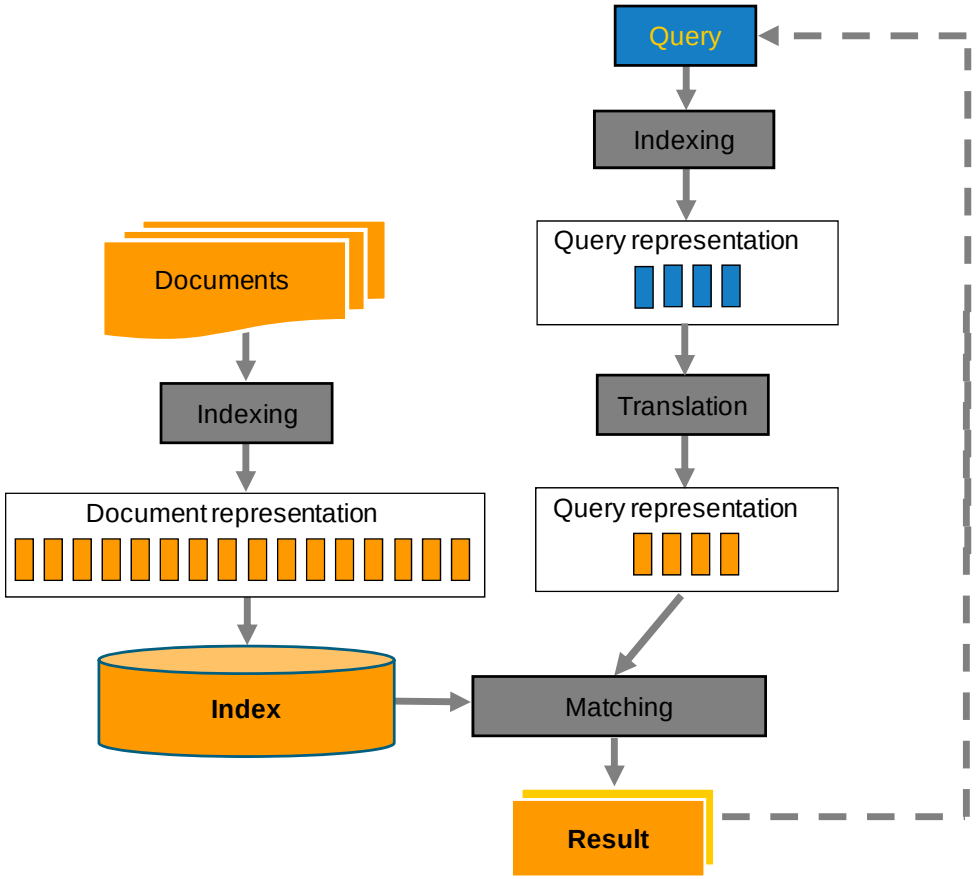

Figure 3: A basic system for bilingual CLIR. The workings are very similar to the monolingual case, with query translation added.

Figure 4 shows a more advanced setup, catering for a document collection containing many different languages. Documents are indexed in all those languages, potentially using language-dependent indexing components (see Section 2.1.6, steps 4, 5, 6). As a result, a set of different index structures, one for each language, is built. Alternatively, a unified index containing documents from all languages can be built, but the system then needs to be able to compute its internal statistics taking this multilinguality into account. For retrieval, the query needs to be translated into all the languages, and a number of matching steps needs to be carried out. The result, a set of ranked lists, is finally merged in order to present the user with one, unified result.

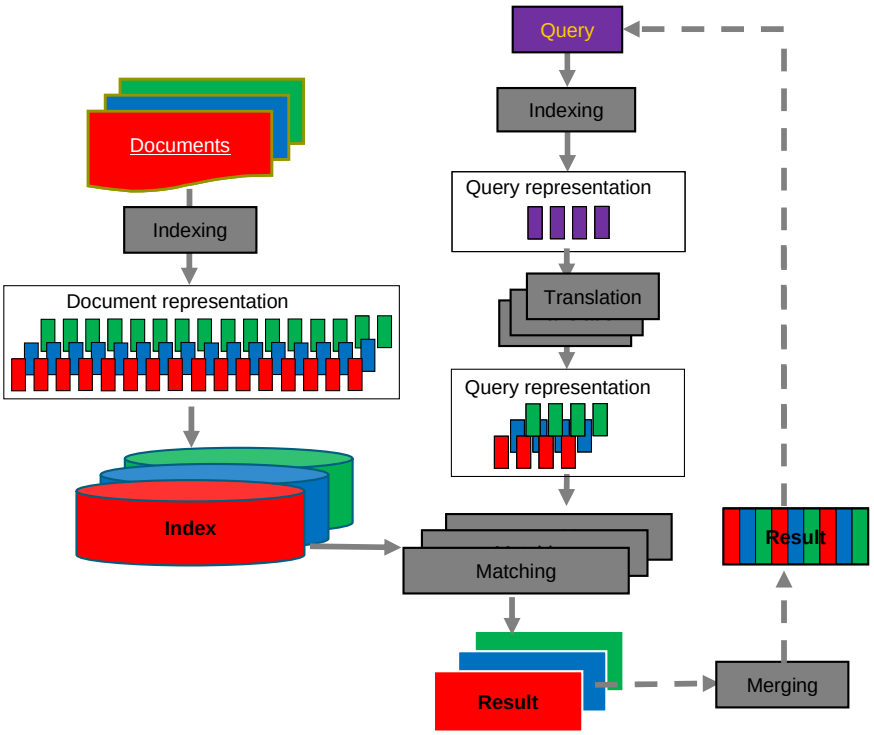

Figure 4: Document translation. Systems can either use a set of monolingual indexes (pictured) or a single, unified index. Retrieval is in the form of a series of bilingual matching steps. A merging step is necessary to unify the result lists.

Figure 5 shows an approach using translation of documents instead of queries. In this approach, all documents are translated into a single language ("interlingua"), potentially a language that is not among the set of languages used in the documents. The query is translated only once, into this interlingua. The match is between the document and query translations in the interlingua, and after back translation, a ranked result list in the user's preferred language is returned.

It is impossible to list all possible further alternatives in a similar way, but the four exemplary flows should provide a good understanding of the basic interaction between indexing, translation and matching components.

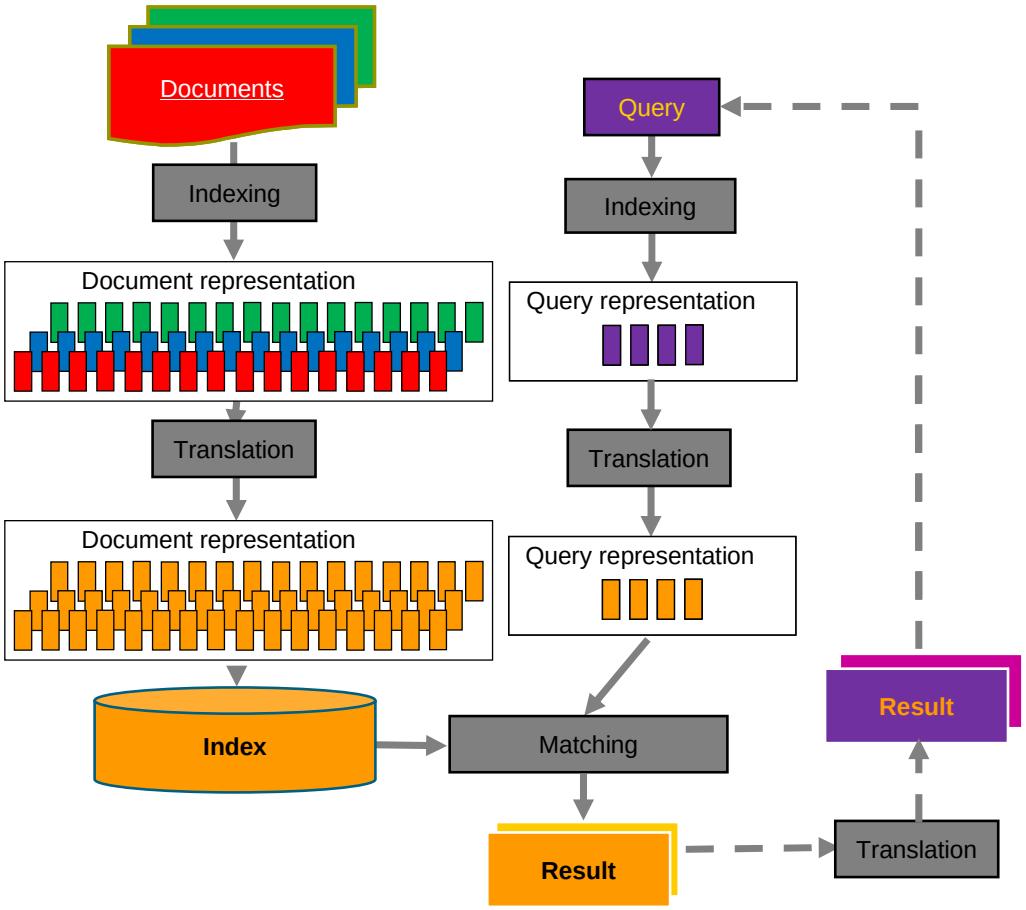

Figure 5: The workings of a system that uses an interlingua for matching. As illustrated, the interlingua need not necessarily be one of the document or query languages. Both documents and queries are translated to the common interlingua.

2.1.5 MLIA/CLIR Blueprint

In (Braschler & Peters 2004b) we have presented a "blueprint" for successful MLIA experiments at CLEF. An analysis had shown that for the specific tasks studied at CLEF, a system designed along the following rough guidelines tended to perform best (all three best performing systems in the 2002 multilingual track had adhered to this same "base formula": (Braschler 2004c), (Chen & Gey 2004), (Savoy 2004)):

- effective, well-tuned monolingual retrieval for as many languages as possible (robust stemming, well-known weighting schemes, pseudo relevance feedback)
- combination of different sources of translation information from different types of translation resources (machine readable dictionaries, machine learning/statistical approaches, machine translation)
- merging of multiple, well-tuned bilingual retrieval results

Our updated analysis conducted for this report indicates that this basic formula still seems valid today. However, while this formula allows participants to score highly in the academic experiments, it remains unclear from the viewpoint of a practitioner if this blueprint generalizes to a concrete operational setting. Thankfully, more information is available today from new CLEF experiments. We will highlight this work when discussing the individual components (Sections 2.1.6, 2.1.7, 2.1.8).

The validity of the blueprint “formula” was also discussed among the attendees of the Winterthur workshop, and it was agreed that it was consistent with the experiences of the practitioners present at that event. We have thus investigated in the following on how to best implement the above “formula”.

Please note that, notwithstanding the similarities in

Figure 2 and Figure 3, it is very hard to add multilingual or cross-language information retrieval on top of existing, monolingual retrieval systems. The overwhelming majority of CLEF experiments use probabilistic retrieval systems that deliver ranked lists of documents in response to user's requests. This is almost to the exclusion of other approaches (such as Boolean retrieval). Ambiguities in translation between languages make probabilistic approaches a very good choice, as term weighting provides a good mechanism to deal with translation uncertainties. It is not possible to adapt most of the approaches originating in CLEF experiments to systems using Boolean retrieval.

2.1.6 Indexing

Indexing components are applied to both documents and queries. Indexing of documents is carried out "offline" – i.e. usually independent from queries. Based on the representation of the documents after a series of indexing steps (generally implemented in the form of indexing components), the index structure proper is built, which is allows efficient access to the documents during retrieval. The most common indexing steps can be summarized as:

1. Format conversion, character conversion, pre-processing
2. Language identification
3. (Document formation)
4. Segmentation, Tokenization, Parsing
5. Feature Normalization
6. (Enrichment) (Entity Recognition, ..)

Individual CLEF experiments do not usually cover all of these steps, most of the focus is generally on steps 4 and 5. Steps 3 and 6 are not covered by most CLEF experiments, and are therefore only briefly summarized in this report. Step 1 and 2 are often implicitely handled through the way that training data is prepared and distributed in CLEF. We will discuss their most important aspects.

Step 1: Format conversion, character conversion, pre-processing

Information retrieval systems nowadays often fill a role analogous to "data integration", by providing users with a single interface to access information from many sources, such as multiple different databases, intranet content, personal files and folders and other data collections ("integration at search time"). Inevitably, inconsistencies in the coding systems used by all these sources arise, especially when multiple languages, potentially using non-Latin character sets, are involved. The general consensus in recent CLEF campaigns is the use of Unicode coding to avoid these problems, and indeed, documents/queries in non-Latin languages have been distributed in Unicode in CLEF from the start. We recommend not deviating from Unicode (or alternatively the ISO-8859 codesets for some European languages) unless there are pressing considerations with integration into legacy systems, especially since Unicode also harmonizes well with XML (see below).

None of the methods or algorithms covered in the following sections make specific use of properties of special file formats (such as PDF or Word) beyond the use of (limited) structural information (such as titles, headings, etc.). This structural information can easily be encoded in XML, which allows good interoperability between different components. We recommend the use of XML for encoding documents and queries during the retrieval process. Please note that this represents no significant restriction in the possibilities available for document presentation to the user – the use of the XML representation of the documents can be restricted to the processing by the retrieval system.

Recommendation	Based on
Use Unicode and potentially XML to encode documents	Successful use in the CLEF campaigns; seems to be able to encode all necessary information

Additional pre-processing, such as removal of non-content bearing sections of certain documents (headers, footers, navigational areas etc.) also usually takes place at this stage.

Step 2: Language Identification

Some of the processing in later indexing steps, such as tokenization, stopword elimination and stemming (see steps 4, 5) is usually language-dependent. In cases where there is no clear indication on the language of a given document or query (e.g. through metadata fields or user profile information), a language identification component needs to be employed. The tasks carried out by the participants in the CLEF campaign do not directly address this issues, although in some of the experiments language identification components are employed. There seems to be no special requirements for language identification that are specific to MLIA/CLIR, and the issue is thus not covered in-depth in this report.

Step 3: Document formation

The tasks in the CLEF campaigns mostly work on document level, with some notable exceptions (e.g. some question answering or Web exercises). The documents that have been used can range considerably in length, from lengthy newspaper articles to shorter newswire documents to very short library records. However, there are scenarios where systems need to operate on sub-document (passage, paragraph, sentence) or super-document (set of documents, folder, linked documents) level. In some cases, such scenarios can be accommodated by splitting/merging documents prior to indexing. If this is not the case, the architecture of the system, and its processing of the documents needs to be adapted. We will not cover these cases, since the experiments we refer to as a base of our analysis exclude this aspect from consideration.

Step 4: Segmentation, Tokenization, Parsing

Before retrieval, documents need to be split into shorter units, in order to allow matching with the query (which typically is either a short natural language description or a set of keywords). For most European languages, the obvious choice is a segmentation into words. Please note, however, that the term "word" has a very specific meaning in linguistics. In information retrieval, the definition of word can be more ambiguous. As we will shortly discuss, it is not always clear what the best "words" from a retrieval perspective are. A more neutral way to discuss these issues is to talk of "terms" or "features", both referring to the units that are ultimately output by the information retrieval system after this step 4. If the experiment is focused on retrieval of text, "term" is often used to describe these units, while "feature" is a good choice if non-textual content is also covered in the discussion (such as multimedia content). Both terms and features occur as a stream of tokens, an ordered list of units output by a segmentation or tokenization component. In step 4 we cover how to produce *valid* units for retrieval, in the sense that the terms or features are usable for subsequent retrieval. Most index structures in information retrieval systems do not allow to match on parts of terms or features – only full, exact matches of terms or features are possible. It is therefore crucial to produce the "right" set of features that leads to a maximum effectiveness during retrieval. This aspect will be covered in more detail in step 5 below.

The easiest option to produce a valid stream of tokens from text written in European languages is to segment using whitespace characters (space, newline, tabulator, etc.). All the characters between two sequences of whitespace characters are treated as a token. Usually, this is not a good option, as special characters such as commas, exclamation marks and others would be retained after such processing. These characters can prevent later matches. An obvious solution to this problem is the restriction of

tokens to sequences of alphanumeric characters. Even such a solution needs careful handling of characters with diacritical marks, however.

While a viable solution for some applications, there may be need for more sophisticated processing if the system needs to be robust especially with regard to named entities. By restricting tokens to alphanumeric characters, a number of issues arise. Potential terms such as "O'Brien", "F/A-18", "Coca-Cola", "Yahoo!Mail" and others are split into multiple tokens. If such splittings areto be avoided, either a dictionary of named entities or more sophisticated linguistic processing is needed. It is very hard to quantify effects such as this splitting of named entities using the Cranfield methodology, and since the issue affects only few queries it is not often addressed in experiment descriptions.

The situation is even more unclear in some Eastern Asian languages, notably Chinese and Japanese. In these languages, no whitespace is usually given between "words", with whole sentences written as continuous strings of characters. Analogous to the treatment of text in European languages, these sentences must be split into units suitable for retrieval. This is not a simple task, as for non-trivial sentences there are often multiple plausible splittings. Literature dealing with the problem usually recommends one of two basic alternatives: the use of word n-Grams (not to be confused with character n-Grams covered in step 5) or the use of a specialized segmentation component. In Chinese, the characters ("ideograms") stand for "basic concepts", and each word in the Western sense is represented by a number of ideograms. Often it is possible to infer the meaning of a word from the meaning of the individual underlying characters. The simplest strategy is therefore to use single Chinese characters as the unit for retrieval, but since there is important additonal meaning encoded in the multi-character words, the performance is usally not optimal. An alternative solution proposed is the use of bigrams of Chinese characters, i.e. overlapping pairs of characters. The use of single characters of retrievals is consequently also called "unigram" indexing. The unigram and bigram strategies can be combined, with the segmentation component outputting a stream of unigrams and overlapping bigrams.

Instead of the use of word n-Grams, there are segmenters available for Chinese. These attempt to find the most probable splitting of a sentence into Chinese words of arbitrary length. Again, as mentioned in the initial remarks on general requirements, we do not necessarily need a linguistically correct segmentation for effective retrieval. It may well be that the more simplistic word n-Gram method allows interesting conflations of related words that would otherwise be represented by longer character strings. These effects are similar to stemming and decompounding for European languages (see step 5).

Abdou and Savoy (2006) compare the use of word n-Grams to the use of a segmenter for Chinese. They conclude that the use of unigram+bigram combination can be competitive with full word-based segmentation. Japanese uses characters loaned from Chinese ("Kanji") as well as two additional syllabaries, namely "Hiragana" and "Katakana" and the Latin alphabet. As in Chinese, sentences are written as continous strings of characters in those four different writing systems. Segmentation can take place in the form of unigrams, bigrams, combination of uni- and bigrams and word-based, again analogous to Chinese. Savoy (2005) reports competitive performance for the combination of unigrams and bigrams compared to full word-based segmentation.

Most information retrieval systems contain a component that removes non-content bearing tokens (also known as "stopwords" or "stop words") from the segmented document stream. Historically, the use of such a component stems from the observation that frequency counts of words are very uneven in languages. Very few words are used extremely often, while most words are used only very rarely (a rule also referred to as "Zipf's law"). When processing the tokens during indexing, a substantial part of the tokens can stem from very few different words. To illustrate this point, consider Figure 6. When processing a full year of articles from German newspaper "Frankfurter Rundschau", roughly 800,000 unique features (word forms) were found. However, the 10 most frequent features make up 16.64% of tokens, and the top 50 most frequent features cover 33.02% of tokens. To reach a 90% coverage of tokens takes roughly 30,000 unique features, meaning that the remaining 10% are split between roughly 770,000 very rare features. Indeed, a clear majority of features (nearly 500,000) occur only once or twice in any of the articles. Most of the highly frequent words (word forms) are articles, particles, conjunctions, prepositions and the like. These fall under our definition of "stopword". It is

argued that most of the meaning of a text is retained even when these words are deleted, and thus historically, in the development of information retrieval systems, they were eliminated primarily in order to reduce the size of the resulting index. We would argue that index size is today often no longer a concern, and stopword elimination purely on grounds of reducing indexing size should be avoided if possible, as we will expand on later.

Apart from reducing the number of tokens to be indexed, stopword elimination has also been shown to be beneficial for retrieval effectiveness as measured by recall and precision. However, we argue that this result should be interpreted carefully. Consider that what is reported in CLEF experiments is often average performance over a number of queries. A small increase in average performance may well hide a performance regression for a substantial number of queries. Furthermore, even though stopwords are often termed "non-content bearing tokens", clearly any elimination of tokens from a text leads to information loss, however small. Often even words such as "the" and "who" can carry critical importance, for example when looking for information on the rock band "The Who". The fact that certain weighting schemes benefit from stopword elimination should therefore be seen more as a deficiency of said weight schemes (which do not properly weigh the stopwords) rather than a proof of the value of stopword elimination to boost retrieval effectiveness. Unfortunately, there is little work published about which weighting schemes are particularly robust towards lack of stopword elimination, but unpublished work by Savoy hints at significant differences (Savoy 2009). We would argue to avoid stopword elimination if possible, alternatively keep the stopword list as short as possible, and carefully watch forthcoming studies on the topic to chose the most robust weighting schemes.

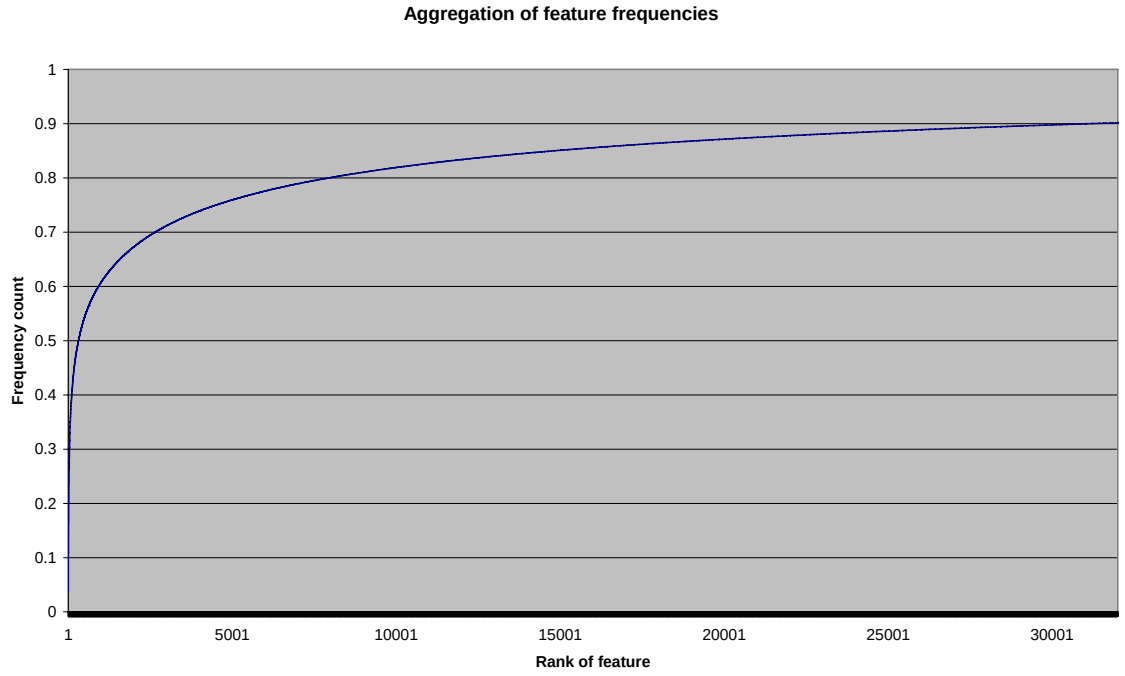

Figure 6: A plot of aggregrated frequencies of tokens. The y-axis gives the ratio of the sum of the frequencies of the top-n ranked features compared to the total frequency of all features (the total count of tokens in the collection). The x-axis represents the 30'000 most frequently encountered features. Few features represent a large part of all tokens in the collection. Example calculated on a collection of documents from the German newspaper "Frankfurter Rundschau".

Recommendation	Based On
Use minimal stopword elimination if possible, choose weighting scheme that is robust with respect to stopwords. This will minimize information loss in phrasal searching	Results by Savoy that show competitive retrieval performance without removing stopwords (Savoy 2009)

Step 5: Feature Normalization

Once valid candidates for retrievable features have been identified, there is an optional step for normalizing these features further, in the hope of enhancing their potential for matching between query and documents. As natural language often allows many different ways of conveying the same information, it is unlikely that the formulations of information need by the queriers would always match the phrasings used by the authors of the documents that contain the relevant information. Differing word surface forms are one of the main reasons for this, as the same basic word can occur in a number of different word forms depending on grammatical gender, number, case etc. There are two basic word formation processes in play: inflection and derivation. Other hindrances to matching include inconsistent use of capitalization (e.g. languages such as English capitalize words at the beginning of a sentence that are otherwise written in lowercase, furthermore, in English, most words are written in with a capital first letter when used in titles) and inconsistent use of diacritics (e.g. in French, diacritics are seldom used when a corresponding character is written in uppercase, also, in languages such as German and French, users often do not write characters with diacritics if they are not easily available on a keyboard, and use corresponding "basic" characters instead).

Nearly all academic information retrieval systems normalize capitalization, usually converting the entire text to lowercase. There seems to be a broad consensus that little is to be gained from keeping the case information, due to the difficulties of handling sentence boundaries and special cases such as titles. Note, however, that some components that do deeper grammatical analysis may depend on case information, which thus has to be preserved prior to the application of such components.

The picture is less clear with respect to the handling of diacritics. There are many reports of CLEF experiments that mention the issue in passing, however, often it is merely stated whether or not diacritical marks are removed – without further analysis on the effect this choice has for subsequent retrieval. An exception to this is work by McNamee and Mayfield (2003), which reports on the changes to retrieval effectiveness when removing or retaining diacritical marks for eight languages – DE, EN, ES, FI, FR, IT, NL, SV. They observe only tiny, insignificant differences based on handling of diacritics. We advise to remove diacritics in order to maximize the potential of matches between queries and documents, unless there is an issue with presenting the so normalized features to the user in an subsequent interactive process.

Recommendation	Based on
Remove diacritical characters from queries and documents. This will minimize mismatches in case of inconsistent use by the querier and authors of documents	(McNamee & Mayfield 2003)

Stemming is an attempt to minimize mismatches between queries and documents due to use of differing word forms.

While it seems intuitive that some form of normalization of word surface forms to common representations ("base forms") should be desirable given the likely query/document mismatch, analysis of the problem shows that more care is needed. While the meaning of a word may often not change between different forms, there are also situations where there is a shift, and where the difference is crucial. In some cases, the querier would explicitly ask for something in the plural, knowing that any occurrences in the singular of the same concept is likely not to lead to relevant information. Normalization of word forms originating from derivation can also be misleading. For example, the word "formation" may be derived from "form", but it is not immediately clear if documents containing the latter word would be helpful in case the former concept is expressed by the querier. In academic literature, the terms "overstemming" (conflation of two word forms that is detrimental to retrieval effectiveness) and "understemming" (failure to conflate two word forms that is detrimental to retrieval effectiveness) are often used. Less well researched are questions of acceptance of stemming by the

user. Stemming often leads to "artificial", truncated representations of words, that are not easily recognizable by the user as desirable for retrieval (e.g. the famous Porter stemmer (Porter 1980) suffers from this phenomenon). While stemmed forms can often be used exclusively for internal representation and matching, there are cases where the user is prompted for interactive feedback on search terms during the retrieval (such as in "relevance feedback", for example), In such cases stemmed representations may not be usable.

Stemming has been shown to be effective for a large number of languages. Braschler and Ripplinger (2004d) give a short overview of results for a number of languages where benefits from stemming have been observed: Slovene, Hebrew, Dutch, German, Italian, French. Very helpful are also the numbers reported by Savoy (2006) on a large number of CLEF collections. Results on effectiveness of stemming in English are not always conclusive, as English has a comparatively simple morphology. While Harman (1991) reported that stemming gives no benefit, Frakes (1992) and Hull (1996) claim at least a small benefit.

Summarizing the reported experiments from CLEF, stemming seems to be beneficial or at least not detrimental on average in all languages, and should probably be used in cases where there is no conflict with presenting stemmed representations of words to users. Depending on the application, it may be beneficial to allow expert users to toggle the stemming function on and off, to allow them to control understemming/overstemming effects. Stemming can be detrimental to retrieval efficiency, which may be a concern in systems with massive processing loads.

Recommendation	Based On
Use stemming during indexing. Potentially allow the user to toggle this feature on/off, if issues of overstemming/understemming are of concern. This will maximize the potential for matches between search terms and documents (thus also aiding in matching)	e.g. (Savoy 2006), (Braschler 2004d), (Hull 1996) and many other CLEF experiments

A problem related to stemming is that of decompounding. A number of languages, such as many Germanic languages (German, Swedish, Dutch), Finnish and Korean, among others, offer a compound formation mechanism whereby the speaker can form new compound words from multiple "basic" words. This compound formation mechanism is not completely alien to English, where a few examples exist (airplane from air+plane, software from soft+ware). However, in English, much in contrast to other languages, this phenomenon occurs too rarely to be of much concern for retrieval. When compounding is an important feature of a language, such as in German, good handling of compounds becomes important for effective retrieval. Braschler and Ripplinger (2004d) show substantial gains by splitting compounds. This is due to a free choice by the speakers of these languages to either use the compound term, or to paraphrase the meaning of the compound term by using multiple words. If the querier and the author of a document decide differently in this matter, mismatches will occur. Unfortunately, there are few systematic studies into different decompounding algorithms. We advise that should decompounding be available for a language with a rich compound formation process, that this option is used, should performance considerations allow it.

Recommendation	Based On
Use decompounding for languages that have productive compound formation, such as German, Dutch, Finnish and others. This will take account of differences in formulation between querier and the authors of documents	(Braschler & Ripplinger 2004d)

We confirmed during our analysis of CLEF papers that the use of stemmers (rule-based word form normalisers) is widespread in the experiments. If no stemmer is available for a specific language, work by McNamee indicates that language-independent character n-Gram techniques are helpful (McNamee 2008) (not to be confused with word n-Grams used to solve the segmentation problem in Eastern Asian languages, see step 4). For character n-Gram retrieval, words are split into sub-units consisting of a set of overlapping strings of characters, typically between 4 and 6 characters long. For example, the word "airport" may be split into character n-Grams of length 4 as follows: "_air", "airp", "irpo", "rpor", "port", "ort_". The technique usually yields a number of common character n-Grams for words that should be conflated, and these tend to give a pretty good representation of the stem of the word. Note, however, that the technique inflates the size of the index, with the actual size depending on implementation and choice of length of n-grams. There is also an issue of acceptability to the user, as unrelated words in queries and documents may match based on common character n-Grams, making it very hard for users to understand why certain irrelevant documents are returned by the system.

Recommendation	Based On
Use character n-Gram techniques in case no other resources for stemming are available. Consider the acceptability of matches based on sub-units of words	(McNamee 2008)

Step 6: Enrichment

A number of additional processing steps can be undertaken before the index is built. These include phrase (multiword) detection, named entity recognition, the use of thesauri to add synonyms and others. These are usually very specific to the concrete setting the system is used in, and the CLEF experiments give little indication on how to generalize. Some of these issues are addressed in the experiments conducted for the question answering track at CLEF.

2.1.7 Translation

Any form of truly multilingual information access (i.e. at least bilingual, involving two distinct languages, but potentially covering an arbitrary number of languages) needs to bridge the language gap between the querier's formulation of information need and the information encoded in different languages in the documents. As mentioned earlier, this "bridging" can take place in essentially four different forms:

1. query translation, i.e. the translation of the formulation of information need
2. document translation, i.e. the translation of the retrievable items
3. both query and document translation, usually by translating both into a common third language, a "pivot" language or "interlingua"
4. no explicit translation, but use of alternative techniques such as sub-word matching or reliance on cognates.

Not all these options are suitable for every multilingual information access scenario. The choice for one of the options over the others is often motivated by the list of languages to be handled and the

resources available for them. In terms of availability of translation resources and picking the right ones for a specific document collection, the academic CLEF papers are of limited value to the practitioner. There are some general "rules of thumb" that can be derived from the CLEF experiments, as we will outline in the following paragraphs. However, it is not possible to compile lists of appropriate translation resources directly from the experiment description. This gap is filled by deliverable 5.2 and related material published on the TrebleCLEF best practice portal.

Usually, for document translation the availability of a machine translation (MT) system is crucial. The use of bilingual dictionaries can often be problematic, as a simple word-by-word translation of documents tends to "blow up" the size of the document due to word sense ambiguities. This happens in a very uneven way, as different words have different degrees of such ambiguity. For example, when translating a three word query between two languages, one of the words may have an unambiguous translation, whereas the other two words may have, say, three and seven alternative translations, respectively. The resulting translated query of eleven words needs careful treatment to avoid giving undue weight to the most ambiguous original query word. For documents, this "expansion effect" exists in an analogous way. Machine translation systems have advanced word sense disambiguation, and attempt to find the best single translation for a given sentence. Statistical translation approaches mostly use probabilities of mappings between terms in multiple languages. They may not be always suitable, as often the number of target terms needs to be specified (i.e. how many terms in the target language are to be produced for a given document), and it is not immediately clear how to select such a number to best represent the document. Translating all documents is a very costly operation, both in computational cost (translating very large collections can easily require days to weeks of computations, making it hard to keep rapidly changing document collections up-to-date) and in storage cost (the documents collection is replicated in the new language). This makes translating the documents to a large number of languages quickly infeasible. If many different language pairs need to be covered, an interlingua (a common intermediate language for both query and document representation) is a possible solution. Potentially, the use of an interlingua needs the combination of two translation steps (query to interlingua and document to interlingua) which is liable to multiply translation issues. However, as results by Savoy (2009) show, the careful choice of an interlingua with good language resources may actually be preferable to a direct translation, when resources for the language pair in question are of bad quality. An advantage of document translation is the avoidance ofextra processing at query time if documents are available in all the query languages.

Query translation scales somewhat better in scenarios with many languages when an interlingua for documents is not an option, or when the computational cost and storage requirements for translating the entire collection are too high. If the system consists of documents in a number of languages, the query needs to be translated into each one of them, potentially a significant performance penalty at query time. Also, resources for all pairs of query language/document language need to be available, unless an interlingua is used for double query translation.

A study by Mandl et al. (2008) demonstrates that multilingual retrieval exhibits much stronger variability in performance from query to query than monolingual retrieval. This means that it is hard to build robust retrieval systems that minimize the number of queries with very bad results (no or only a small fraction of relevant items retrieved). A careful investigation of translation accuracy query by query for a representative query sample is highly recommended. We assume that systems will benefit greatly if translation can be adapted to cover domain-specific vocabulary and minimize "out-of-vocabulary" problems. This assumption was underscored by the experts that attended the Winterthur workshop (Braschler & Clough 2008).

A small study by Savoy (2009) into the most frequent translation issues in a machine translation system (Google MT translation) shows that polysemy is the leading factor in bad translation for European languages (German, French, Spanish). Again, this problem can be mitigated by using domain-specific resources, if available. There is limited research into the right choice of translation resources within CLEF, as participants often use what is available to them. However, more insights were given during the discussions at the Winterthur workshop, where the experts generally agreed that

maximizing coverage (to solve the out-of-vocabulary issue) and integrating as many sources for translation as possible (specifically also domain-specific or enterprise-specific ones) is key to successful multilingual retrieval in operational situations.

This form of combination of translation resources can be extended to a combination of different types of translation resources. It is possible to e.g. combine machine translation with lookup in machine-readable dictionaries, in order to boost the vocabulary coverage. Likewise, statistical approaches can be combined with MT and machine-readable dictionaries. It has been shown that this form of combination can boost robustness of the system. Of course, using such combination approaches, both cost for acquisition of resources as well as computational costs will increase. See also (Braschler 2004c), (Savoy 2003), (Savoy 2004).

A possible alternative remedy to OOV problems is the use of pre-translation expansion. For this technique, the query is expanded with additional, related terms by using a pseudo-relevance feedback step (see also Section 2.1.8 for a similar technique prior to retrieval). A "pilot" collection in the source language is used for an initial search, yielding a ranked list of documents matching the source query. From the top ranked documents in this list, additional characteristic terminology is extracted and tacked onto the query. The hope is that a longer query is less liable to suffer from missing translation for some of its terms. Prerequisite for the use of this technique is a suitable pilot collection for all the source languages to be covered. See also (Ballesteros & Croft 1997) for a description of the technique.

There are some studies about using no explicit translation step for cross-language information retrieval, but in contrast to language-independent approaches for monolingual retrieval, their application is much more limited. Buckley et al. (1997) present work where they treated English as "misspelled French", matching on cognates between the two languages. This work may have potential to be combined with pre-translation and post-translation expansion. It is unlikely that the approach translates well to language pairs where there is little overlap in vocabulary. Likewise, Gey (2005) has demonstrated retrieval on Japanese documents using Chinese topics, relying on commonalities in the use of Kanji characters between the two languages. It is unclear how this work could be extended to other language pairs. Since such "resource-light" approaches are especially interesting for cases where appropriate language resources are hard to locate, the limitation on few specific language pairs is probably a critical one.

Translation resources are often only bilingual. If query translation is used to access a collection of documents written in more than two languages, a set of consecutive bilingual translations and retrieval steps is necessary. The output of these steps will be a set of result lists, each monolingual and for a different language. In order to present a single, integrated result list to the user, these monolingual lists need to be merged.

Merging remains one of the bigger issues in multilingual information retrieval. The issue has been extensively looked at in the CLEF multilingual retrieval task (Braschler 2004a), but remains as of today largely unsolved. The problem stems from the distribution of relevant items in the multiple lists to be merged: all popular weighting schemes used to calculate the retrieval scores for documents return values that are only valid for relative comparison, i.e. indicating a higher probability of relevance for a document A if its score exceeds that of document B. The absolute values of the scores, however, do not lend themselves to easy interpretation, since while they are a function of the probabilities of relevance, they do not represent the probabilities themselves. It is therefore not possible to predict the number of relevant items contributed to an overall merged result by the individual monolingual lists in an easy way. As we cannot assume that relevant items are necessarily distributed evenly across languages, this presents a problem: an algorithm is necessary that dictates on how many items to pick from each monolingual list in turn, and how to rank them overall. Some possible merging methods are described in Savoy (2004).

As is shown in (Braschler 2004c), it is possible to retroactively calculate an optimal merging for a training collection based on human assessments of relevance. Based on this "gold standard" it can be shown that all simple approaches to merging, such as "raw score merging", "interleaving", and their normalized counterparts, substantially underperform the optimal merging strategy by up to around

40%. This substantial performance degradation can be an incentive to chose document translation, should the computational and storage cost incurred be acceptable.

Recommendation	Based on
Maximize coverage of translation resource; add domain-specific resources. This will minimize the number of queries that fail due to missing/incomplete translations	Winterthur Workshop (Braschler & Clough 2008)
Using document translation solves the merging problem (if computational cost acceptable)	(Braschler 2004c)
Combination of different types of translation resources (if acceptable considering availability, computational cost and financial cost)	(Braschler 2004c), (Savoy 2003), (Savoy 2004)
Interlingua can help in cases where direct translation resources have questionable quality.	(Savoy 2009)

2.1.8 Matching

The analysis of the overview papers of past CLEF campaigns and of descriptions of experiments by participants underscores that good monolingual matching in all languages to be covered is a prerequisite for effective multilingual retrieval. Experiments by Dolamic et al. (2008) and McNamee (2008) give an excellent indication on how to proceed. CLEF has addressed different kinds of text over the years. Notably, while starting out in 2000 with mostly newspaper text, CLEF has studied in 2008 how approaches translate to very short documents (specifically, bibliographic records). A number of weighting schemes have proven to be consistently close to best performance for a variety of these tasks. Among these are Okapi/BM25 (the most used weighting scheme in CLEF over the years), language modeling (LM) approaches and deviation from randomness (DFR) (see (Dolamic et al. 2008) for all three weighting schemes), as well as lnu.ltn (see (Singhal et al. 1996)). Performance differences between these approaches are usually not statistically significant, with the two newer approaches LM and DFR scoring a bit better in Dolamic et al.'s most recent experiments. All these approaches score well above simpler baselines such as tf.idf – see e.g. (Schäuble 1997).

Recommendation	Based on
Use one of the consistently high-performing weighting schemes such as Okapi/BM25, LM, DFR or lnu.ltn	(Dolamic et al. 2008)

Monolingual retrieval performance can benefit from pseudo relevance feedback techniques. This is a technique that addresses issues in verbalization of an information need by the user (see also Section 2.1.1). For pseudo relevance feedback, the system extracts characteristic terms from the top ranked documents and uses them in an enhanced query for a second retrieval step. Again, our survey of CLEF papers has shown wide-spread adoption (over 100 different experiment descriptions mentioning the technique). However, pseudo relevance feedback entails a substantial performance penalty, requiring at least a duplication of processing time. This should be considered when deploying in operational systems. Both stemming and pseudo relevance feedback lead to matches between documents and queries that do not necessarily contain the search terms as entered by the user. While highly beneficial for boosting recall, the acceptability of this needs to be carefully considered based on system usage. Overall impact on average effectiveness is mixed, as is reported in e.g. (Dolamic et al. 2008) and (Moulinier & Williams 2005). This makes the technique mostly interesting in scenarios calling for high recall, or as an optional processing step that can be toggled on and off by the user.

Recommendation	Based on
Use pseudo-relevance feedback as an option to boost recall. Extra terms from the enhanced queries may lead to matches that do not necessarily contain the search terms as entered by the user. Acceptability of this phenomenon needs to be considered	Wide-spread use by CLEF participants, good results in many experiments – see e.g. (Dolamic et al. 2008) and (Moulinier & Williams 2005).

3. User-oriented Multilingual Information Access

The interface acts as the intermediary between users of information retrieval (IR) systems and the search system itself. A well-designed interface should assist users in clarifying their information needs, and subsequently help them formulate suitable queries and understand the results. However, interactive multilingual information access (MLIA) systems provide an additional challenge to designers, because users may not have the necessary language skills to find and interact with documents written in multiple languages. To provide effective access to multilingual document collections, users require search assistance. In this section we summarize best practices to support interaction at various stages within the search process. We will use two sources of evidence:

(i) **Interactive Cross-Language IR experiments**, mainly in the framework of CLEF. As explained earlier, we will focus on the results of CLEF because it is the only major evaluation campaign that is primarily focused on Cross-Language aspects of Information Access, and in fact is the only one that has run an interactive multilingual information access track (iCLEF) for many years (2001-2006 and 2008-2009).

(ii) **Best Practice Recommendations**, mainly taken from a report which is the outcome of a workshop bringing together researchers and representatives from current and potential user communities for Multilingual Information Access applications.

3.1 Interactive Cross-Language Information Retrieval Experiments

The Cross-Language Evaluation Forum (CLEF) has been devoted to the study of Multilingual Information Access problems since its foundation in 2000. Since 2001, the interactive track, iCLEF[12] , has been focused on the problem of Multilingual Search *assistance*. In seven years it has addressed two main aspects of the problem: (i) document selection and results exploration; and (ii) query formulation, refinement and translation. Both aspects have been addressed for various Information Access tasks (document retrieval, image retrieval, question answering...), from different methodological perspectives (hypothesis-driven, observational studies) and for different language profiles (i.e. different degrees of familiarity of the user with the target language/s).

In this section we summarize major findings in the iCLEF track; we start with two subsections discussing major findings for cross-language document retrieval assistance in the areas of document selection and query formulation. Then we discuss two related MLIA tasks: cross-language image retrieval and question answering.

3.1.1 Document selection and results exploration

A multilingual IR system must provide useable summaries for users to make informed decisions about relevance. A default solution is using an off-the-shelf Machine Translation (MT) system, but this is not necessarily the best solution. First, because full document translation is usually noisy and confusing to users and therefore some kind of cross-language summary might be helpful. Second, because relevance feedback on the output of an MT system is tricky, as there is no direct alignment between the expressions in the translation seen by the user and the actual words in the original documents. Third, because MT systems are not available for all language pairs, and when they are available they have limited performance on domain-specific collections.

Document selection was the focus of the first iCLEF campaign in 2001 (Oard & Gonzalo 2002). To support manual selection in cross-language applications, a translated indicative surrogate for the document must be created. “Indicative”, as opposed to “informative” is used to emphasize that the surrogate is designed to provide the information that a reader would need to decide whether to read the document, rather than directly providing some of the information that the reader might be seeking. Three factors affect the utility of translation technology for the document selection task: *accuracy* – to what extent a translation reflects the intent of the author -, *fluency* – the degree to which a translation

[12] nlp.uned.es/iCLEF

can be quickly used to achieve document selection – and *focus* – the degree to which the reader's attention can be focused on the portions of a translated document that best support document selection.

Before iCLEF 2001, the vast majority of MLIA research had focused on the automatic components of a system, and only a few results had been reported by individual search teams: Oard and Resnik (1999) found that users were able to categorize automatically translated documents more consistently than an automatic classifier, but less consistently than a comparable set of users were able to do when using more fluent human-prepared translations. Ogden and Davis (2000) performed quantitative user studies of cross-language document selection using Systran translations of German documents retrieved by an automatic system, finding that a single searcher with no self-reported German reading skills could identify relevant documents with 99% precision and 86% recall (within the normal range of inter-annotator agreement), which suggests that MT technology might be well suited for the document selection task. They also experimented with language-independent document thumbnail visualizations with colour-coded highlighting of query terms, and found that twice as many documents could be assessed in a fixed time without a significant loss in precision. This is an indication that this kind of visualization techniques might be of special interest in cross-language search. Finally, Suzuki et al (2001) compared document selection on word-by-word translations with translated summaries, finding that their particular choice of translated summaries did not work as well as simple word-by-word translations. However, they adopted a between-subjects design (different people judged relevance of the same document under the two different conditions), which makes a direct comparison of both alternatives less significant. Indeed, we will see that results obtained in iCLEF were very different.

In iCLEF, the evaluation methodology adopted was hypothesis-driven: each research group had to formulate a hypothesis about some aspect of interactive cross-language document selection, design two alternative systems supporting document selection (a *reference* and a *contrastive* system). A within-subjects quantitative user study design was chosen to compare selection effectiveness with different document surrogates, because a within subjects design offers greater statistical power than a between-subjects design. Research groups had to recruit and train as many users as possible (in groups of eight), and run a number of search sessions with a prescribed combination of topic/user/system to filter out topic and user effects and detect system effects (that should confirm or discard the research hypothesis). Three experiments were carried out and later summarized in (Oard et. al 2004). The most significant differences found were:

- A direct comparison between simple term-by-term translation (without any disambiguation) and Systran revealed a statistically significant advantage for Systran full MT versions of the original documents; also, many more documents were judged as "unsure" when inspecting the word by word translations.
- Another direct comparison was made between Systran translations and cross-language document pseudo-summaries consisting of a translated list of noun phrases appearing in the document; translation was performed without MT machinery, using raw statistical evidence extracted from comparable corpora. The result was that, while precision with both alternatives was similar, users were able to judge relevance much faster (52% greater recall) with the translated noun phrases. The difference was statistically significant and corroborated by an observational study of the search sessions and the subjective impressions of the recruited users, collected in pre- and post-search questionnaires.

These results suggest that

- If feasible, producing high quality translations for the documents in the collection pays off: basic term by term translation is not enough to facilitate effective cross-language document selection, and should be avoided unless full MT is not available or viable.
- MT, however, is not the best choice for document selection; high quality, summarized information may lead to similar precision and faster – i.e. easier –decisions on relevance.

The conclusion that summarized translations may have a positive impact on document selection has been confirmed in other iCLEF experiments. In (Llopis et al. 2003), relevance assessments made on document passages (selected by a passage retrieval system) gave better results than assessments on machine translations of the full document. In a subsequent experiment, Navarro et al. (2004) concluded that syntactic-semantic patterns extracted from the passages, containing only key concepts, led to faster relevance assessments with a similar accuracy. Therefore, all evidence supports the use of summarized versions of the document to minimize the cognitive effort of document selection without an adverse effect on precision.

Finally, in (Ostenero et al. 2004) it was shown that cross-language summaries (using reliable translations for noun phrases in the documents in the case of their experiment) could also be used as document representations for the indexing and retrieval steps, and not only as a way of informing users about the document contents. The size of the cross-language summaries was at least three times smaller than the original document; that means that one collection can be indexed in three additional languages at a similar cost tohaving one additional language with full machine translation. The benefits of having the whole collection translated at indexing time are huge from the point of view of interactive retrieval, because it maps the problem into the – much better known – monolingual equivalent of the problem (Oard 2009).

Outside of iCLEF experiments, (Richardson 2007) used concept maps to represent the content of the documents as an indicative summary, also with positive results.

Not all compression techniques, however, give good results. Dorr et al. (2004) report about a technique to compress news article headlines which leads to faster relevance assessment, but less accurate.

Recommendation	Based On
For cross-language document selection, offering high-quality translations to the users pays off: word by word gist translations perform substantially poorer than full machine translation.	(Oard et al. 2004)
For cross-language document selection, cross-language document summaries (noun phrases, relevant passages, key conceptual relations, concept maps) can lead to faster relevance assessments without losing precision.	[Oard et al. 2004), (Llopis et al. 2003), (Navarro et al. 2004), (Richardson 2007)
If possible, translating the whole document collection at index time pays off, because it maps query reformulation and relevance feedback issues into their monolingual (much simpler) version. Using (appropriate) cross-language summaries can be an optimal solution, because they perform nearly as well as full documents, indexes take much less disk space, and are optimal for cross-language relevance assessment.	(Ostenero et al. 2004), (Oard 2009)

In general, the experiments above were run with users having no knowledge of the target language. This is, obviously, the cross-language retrieval scenario that requires more assistance for users. However, experiments at SICS suggested that document selection is also problematic when users have active language skills in the target language. They made two experiments with Swedish users searching English documents and having reasonably good English skills. The experiment reported in (Karlgren and Hansen 2003) suggested that relevance assessment in a foreign language takes more time and is prone to errors, compared to assessment in the reader's first language. In their second experiment (Karlgren 2004) they observed that users discarded bookmarked documents more often in

English than in Swedish, suggesting that the user's confidence in his relevance assessments is lower in a second language.

3.1.2 Query formulation and translation

Query formulation, reformulation and translation are particularly challenging for translingual information access systems. Some of the main issues are:

- How should systems interact with users to achieve optimal query translation? While in a monolingual system displaying search suggestions is just an option, a translingual system must offer the possibility of changing the term translations chosen by the system, and must help the user select the most appropriate translations (at least when the user does not have skills in the target language).
- How should systems manage relevance feedback? This is a particularly tricky issue. Imagine, for instance, an English speaker searching Japanese documents. When he indicates a relevant term in the English machine-translated version of a Japanese document, what should the system do? If MT is a black box in the system, it is not clear how to trace back to the original Japanese term.
- How should systems manage assisted query translation when there are many target languages? The more target languages, the quicker assisted translation can become a mess for a user without language skills in all the target languages.

In the Encyclopedia of Library and Information Sciences, Douglas W. Oard points out that "*whether people can learn to formulate effective queries is at this point the [question about MLIA] we know the least about*" (Oard 2009). In the context of iCLEF, however, we have found some evidence that can be of help for the task of designing MLIA systems.

In the CLEF 2002 and 2003 interactive tracks, research groups interested in the design of systems to support interactive Cross-Language Retrieval used a shared experiment design to explore aspects of that problem. Participating teams each compared two systems, both supporting a full retrieval task where users had to select relevant documents given a (native language) topic and a (foreign language) document collection. The two systems being compared at each site should differ in (at least) one of these aspects: a) support for document selection (how the system describes the content of a document written in a foreign language), b) support for query translation (how the system interacts with the user in order to obtain an optimal translation of the query), and c) support for query refinement (how the system helps the user refine their query based on previous search results).

One of the most basic outcomes of iCLEF experiments is that support for user-assisted translation of the query improves search results. The University of Maryland group (He et al. 2003) compared fully automatic query translation with assistance in form of other terms with the same translation (potential synonyms) and sentences in which the word is used in a translation-appropriate context. Users obtain better results when they use the assisted-query features. In a similar experiment (Dorr et al. 2004), they test supported query translation using two techniques for generating keyword in context examples of usage. Again, users reach relevant documents faster than without assisted translation capabilities.

Recommendation	Based On
User-assisted query translation facilities have a positive impact on search effectiveness, and should be included as a system feature. The more context, the better: definitions and examples taken from corpora reduce the cognitive load and therefore improve the user experience.	(He et al. 2003), (Dorr et al. 2004)

But the fact that user-assisted translation improves search results does not imply that this feature must be shown to the user by default. On the contrary, the results of most observational studies in iCLEF

indicate that this can be annoying to users. Following the minimal cognitive effort principle, users are only interested in checking or modifying the system's query translation when things go wrong. Petrelli et al. (2003) specifically addresses this question (*Should the user check the query translation first*?) and the answer is a clear no.

Recommendation	Based On
In general, users are not comfortable choosing translations for their query terms in a foreign language; this task requires a high cognitive effort. Therefore, query translation should be hidden to the user by default. However, the translation process is usually noisy and leads to irrelevant results; in those cases, the system should be able to explain how query translation was performed, and to improve the translation with the help of the user.	(Petrelli et al. 2003) and most observational studies at iCLEF.

There seems to be, however, intermediate solutions between assisted query translation and full automatic translation that lead to better search results without imposing too much extra effort on the user. López-Ostenero et al. (2005a) present an approach in which the user is asked to reformulate the query prior to search without presenting any term in the target language. They build a dictionary of aligned noun phrases between the source and target languages, taking such noun phrases from the collection. When the user types a query, the system displays noun phrases that seemed to be related to the query and can be translated with high accuracy using the aligned phrases resource. Retrieved documents are then shown as a list of relevant noun-phrases, which are clickable for direct relevance feedback. This system compares favourably to conventional user-assisted query translation, giving 65% better results in terms of Van Rijsbergen's F measure. From the observational study, it seems that users are more comfortable selecting appropriate noun phrases than inspecting foreign-language terms, leading to optimal search results with low cognitive effort.

Recommendation	Based On
Feedback mechanisms that help translating a query better without showing foreign language terms to the user can be effective to improve search effectiveness without adverse effects on the perceived difficulty of the task.	(López-Ostenero et al. 2005a), feedback via suggestion of noun phrases related to the query and taken from the document collection.

An additional conclusion from that experiment is that the document translation and query translation/formulation/refinement facilities must be designed together as a whole to produce an optimal translingual search assistant. To give a negative example, think of using off-the-shelf Machine Translation to display the contents of the retrieved documents when the cross-language search mechanism is assisted query translation; this is a problematic strategy, because it will be challenging to link the outcome of the MT system with the translation alternatives offered by the query translation machinery.

Recommendation	Based On
Document translation and query translation/formulation/refinement facilities must be consistently designed to fit together.	(López-Ostenero et al. 2005a)

3.1.3 Image Retrieval

In the Interactive Image CLEF 2004 (Clough et al. 2005), interactive retrieval of images annotated in foreign languages was studied using a known-item retrieval task ("stuff I've seen before") and the same methodology used in previous iCLEF tracks. Two different research groups submitted results: (Cheng et al., 2005) successfully combined traditional text-based searches with content-based information retrieval, by allowing users to select the dominant color of the target image. This feature was shown to improve effectiveness and also to reduce total search time. (Bansal et al., 2005) tested the usefulness of suggesting up to 10 related terms for the user's query, but the advantages of this functionality could not be proved.

The task was run again as part of the Image CLEF 2005 track. Petrelli & Clough (2006) tested an alternative visualization of the search results. The proposal was able to cluster the results into a hierarchy of text concepts. In spite of the fact that users claimed to prefer this alternative visualization, the results showed that their performance was slightly more effective (in terms of number of target image successfully found) and more efficient (in terms of average time used) when using the simplest system. Remarkably, user perceptions do not always correlate to actual performance as measured extrinsically.

iCLEF 2006 focused entirely on image search on Flickr (a naturally multilingual image collection with millions of photographs tagged and described collectively by world-wide users in several languages). Three different tasks were proposed, with the goal of analyzing different aspects of the image retrieval task from a cross-language viewpoint (Karlgren et al., 2007):

- topical ad-hoc retrieval (recall): Find as many European parliaments as possible
- creative open-ended retrieval: Find three illustrations for an article about saffron cultivation in Italy and cuisine.
- visually-oriented task: What's the name of the beach where this crab is lying?

Artiles et al. (2007) analyzed the behaviour of 22 users with different language skills using a cross-language image search system featuring three search modes: monolingual search, automatic query translation, and assisted query translation. Results showed that even in the most favorable setting (in a task of this kind, results can be judged as relevant visually), users try not to translate into unknown languages unless it is strictly unavoidable. For instance, the image of the crab was annotated in German, but this was not know by the Spanish users; thus they avoided German as a target language until they have given up looking for the image in their most familiar languages (Spanish, English, Italian).

The learning curve of the proposed cross-language facilities was fast (most users were using multilingual translations), although they rarely interacted with the system to fix or improve translations. Similarly, in (Clough et al., 2007) it was shown that users, while having a positive impression of advanced cross-language search facilities, still preferred monolingual searches, whenever they were able to choose. Finally, both experiments showed that cross-language image retrieval was a feasible task, as users were overall able to search effectively.

Recommendation	Based On
Combining text-based (even simple approaches) with content-based facilities can lead to better search effectiveness in less time.	(Cheng et al., 2005)
Advanced organization features (such as hierarchical clustering of search results) can be appreciated by users even if they do not lead to improved search effectiveness.	(Petrelli & Clough, 2006)
Even in an image retrieval setting, cross-language search should not be offered by default, as users have a strong initial reluctance to search in unknown languages.	(Artiles et al., 2007)

3.1.4 Question Answering

The iCLEF 2004 and 2005 tracks included an interactive cross-language Question Answering task which used the collection, questions and assessment procedure of the CLEF QA track. This provided an initial extrinsic evaluation of the quality of the interactive MLIA systems, measuring the user's ability to answer factual questions when searching a collection of foreign-language texts. Overall, the task was found to be feasible but not trivial: factual questions were answered correctly about 50% of the times (the other 50% being unanswered or incorrectly answered).

The experiment methodology was the same as in previous iCLEF tasks; for the assessment of the answers (which had to be manually written by the users), the same evaluation metrics for the automatic task (strict and lenient accuracy) were used to facilitate comparisons. However, this decision lead to the lack of an appropriate evaluation methodology for the interactive task (López-Ostenero et al., 2008). These evaluation metrics cannot capture specific interactive aspects that are of special interest in this kind of task. For instance: did the user fail because she read an incorrect machine translation of the source document containing the answer, or because she needed more context, or because she did not understood the evaluation rules and gave a too-broad answer, or...?

Research groups tested different approaches related to the size of the context shown to users in order to help them find the correct answer. Some groups tested the usefulness of showing the results grouped under ontological concepts or linguistic patterns (Navarro et al., 2005), showing different-size summaries (He et al., 2005) or passages (Navarro et al., 2006), allowing the access to the whole document (Figuerola et al., 2005, López-Ostenero et al., 2008) or filtering the results removing those passages without potential answers type (López-Ostenero et al., 2005b). The general lessons learned after the experience is that the more context is available, the easier it is to find the answer. Other functionalities such as the usage of dictionaries and machine translation facilities (Zazo et al., 2006), and highlighting expressions likely to contain the target answer also appeared to be useful (Navarro et al., 2006; López-Ostenero et al., 2005b; Peinado et al., 2006), at least in terms of user's satisfaction.

Surprisingly, systems featuring more complex approaches performed, in general, slightly worse than simpler approaches, even when advanced features were perceived as useful by the users.

Recommendation	Based On
When user interaction is possible, simple systems may suffice for CL-QA tasks	(López-Ostenero et al., 2005b; Peinado et al. 2006; López-Ostenero et al., 2008)
In QA systems, especially in cross-language approaches involving translation, users need more context (than in a monolingual setting) to avoid finding incorrect answers.	(He et al., 2005; Figuerola et al., 2005; López-Ostenero et al., 2005b; Navarro et al., 2005)
Generally, the best performing setup for CL QA is a standard document retrieval system performing monolingual searches over a translated collection.	(Gonzalo & Oard, 2005; López-Ostenero et al., 2005b; López-Ostenero et al., 2008)

3.2 Best Practice Recommendations from Users' Experience

We will summarize here the result of the workshop entitled "Workshop on Best Practices for the Development of Multilingual Information Access Systems: the User Perspective" held 24-25 June 2008 in Segovia, Spain.

Despite the large and active research community working on Multilingual Information Access (MLIA) topics, and the maturity of many MLIA technologies, there is still a general lack of commercial MLIA systems available, and a lack of adoption of MLIA technologies in user communities with multilingual information access needs. One of the identified problems is the mutual lack of awareness between MLIA researchers and current – or potential - user communities.

In June 2008, the TrebleCLEF EU coordination action organized a workshop entitled "Workshop on Best Practices for the Development of Multilingual Information Access Systems: the User Perspective" (Gonzalo et al. 2009). The workshop brought together MLIA researchers and representatives from relevant user communities, namely: Cultural Heritage, European government agencies, news agencies, patent and trademark professionals, enterprise and web search companies, and EU projects. Its duration was one and a half days, of which half a day was reserved for an intensive working session, where a consensus reaching strategy ("grid of groups") was implemented to reach a common vision on two specific issues: (i) features that MLIA systems should have from the users' perspective and (ii) strategy to provide MLIA technology with these features and transfer these technologies to society, considering the use of evaluation forums such as CLEF.

The input from the user communities was very valuable: in particular, a list of desirable features for MLIA systems emerged as a consensus of the workshop participants, covering aspects such as integration (of cross-language search capabilities with global information access and knowledge management environments), search interface, results presentation and personalization. Here we summarize the main results.

There was general agreement on the fact that a use case must be specified before making an exhaustive list of features: the starting point is a model of a user and a model of the task. For instance, there is little in common between a monolingual web surfer and a patent retrieval specialist with passive knowledge of six languages. However, it was still possible to make some practical observations at a very general level:

Integration

- Systems must be transparent regarding cross-language search. There is no such thing as a Cross-Language Information need: there are simply information needs that cannot be fully satisfied without finding information in other languages. So, by default, there is no need to know what the system is doing and how it is working (but still full control should be available if required, see *search interface* below). Note that this is consistent with the results reported at iCLEF.
- Multilingualism is just a feature of Information Access systems, and it must be seamlessly integrated.

Search interface

- There should be an advanced search mode that gives user full control over multilingual features (target languages, query translations) for the small percentage of advanced users that want control when things go wrong. There were, however, some disagreements between participants as to whether this still holds when query expansions made by the system are very complex. (Again, this is consistent with the outcome of iCLEF experiments)
- If possible, link structured sources that help mapping the meaning of the query (e.g. with named entities). One common suggestion that had not been tried in the iCLEF experiments was considering Wikipedia as a source to contextualize alternative translations.

Results presentation

- Interfaces should be flexible about how to organize results. There must be at least two choices: separated by target language, or merged. The default view depends on the application (and the user profile).
- There should be a choice of seeing the original document or a translation. If translation for a certain language pair is not available, one option is to show metadata: named entities, categories, etc. Another option is to translate into the language which is more familiar to the user (according to his/her profile), or perform some approximate translation (summarization, key concept translation, word-by-word in the worst case) if possible. (Note that in iCLEF experiments certain forms of cross-language summaries were found to be preferable to full document translation).
- When few monolingual results are available, the system should alert the user whenever there is more information available in other target languages.
- The system should warn about the quality of Machine Translation and about how authoritative the translation is, to avoid wrong expectations from the user.

Personalization

- There must be some support to specify language skills and translation preferences in the user profile, and the interface should adapt to this profile. For instance, default translation should only be provided for languages unknown to the user. If a user profile is not available, the system should have clever default behaviour, e.g. not translating Portuguese documents by default if the query is in Spanish (because Spanish speakers have passive, reading abilities in Portuguese).
- Ideally there should not be an upfront log profile. It is better to start with a sensible default, and allow changes/updates.

Note that these recommendations are compatible but, at the same time, different in nature from the outcome of laboratory experiences. Altogether they provide a comprehensive set of best practices to build cross-language search assistants, which is one of the main steps towards Information Access without language barriers. It is only recently that translingual information access has reached the public at large – with translingual search now being offered by major web search engines such as Google and Yahoo –, but these services are still quite primitive in terms of their search assistance capabilities, something that is much more crucial in a multilingual search environment than in a standard monolingual setting.

4. Summary of recommendations

As a wrap-up, we now list all best practice recommendations discussed in this report, keeping two separate lists: one for ad-hoc retrieval components, and one for user-oriented aspects of MLIA systems.

Recommendations (for retrieval component)	Based on
General Requirements	
Use a retrieval system supporting term weighting and ranked retrieval	Necessary pre-condition for most of the state-of-the-art CLIR/MLIA components
Indexing	
Use Unicode and potentially XML to encode documents	Successful use in the CLEF campaigns; seems to be able to encode all necessary information
Use minimal stopword elimination if possible, choose a weighting scheme that is robust with respect to stopwords	Results by Savoy that show competitive retrieval performance without removing stopwords (Savoy 2009)
Remove diacritical characters from queries and documents	(McNamee & Mayfield 2003)
Use stemming during indexing. Potentially allow the user to toggle this feature on/off, if issues of overstemming/understemming are of concern	e.g. (Savoy 2006), (Braschler 2004d), (Hull 1996) and many other CLEF experiments
Use decompounding for languages that have productive compound formation, such as German, Dutch, Finnish and others	(Braschler & Ripplinger 2004d)
Use character n-Gram techniques in case no other resources for stemming are available. Consider the acceptability of matches based on sub-units of words	(McNamee 2008)
Translation	
Maximize coverage of translation resource; add domain-specific resources	Winterthur Workshop (Braschler & Clough 2008)
Using document translation solves the merging problem (if computational cost acceptable)	(Braschler 2004c)
Combination of different types of translation resources (if computational and financial cost are cceptable)	(Braschler 2004c), (Savoy 2003), (Savoy 2004)
Interlingua can help in cases where direct translation resources have questionable quality	(Savoy 2009)
Matching	
Use one of the consistently high-performing weighting schemes such as Okapi/BM25, LM, DFR or lnu.ltn	(Dolamic et al. 2008)
Use pseudo-relevance feedback as an option to boost recall	Wide-spread use by CLEF participants, good results in many experiments – see e.g. (Dolamic et al. 2008) and (Moulinier & Williams 2005).

Recommendations (for user interface)	Based On
Cross-Language Document Selection	
Try to offer high-quality translations to the user.	(Oard et al. 2004)
Cross-language document summaries are often preferable to full machine translation of docs.	(Oard et al. 2004), (Llopis et al. 2003), (Navarro et al. 2004), (Richardson 2007)
If possible, translating the whole document collection at index time pays off. Translating a document summary can also work.	(Ostenero et al. 2004), (Oard 2009), Treble CLEF User Communities Workshop (Gonzalo et al. 2009)
Interfaces should be flexible about how to organize results (by target language or merged).	Treble CLEF User Communities Workshop (Gonzalo et al. 2009)
If translation is not available, try to show metadata, translate into the user's second language, or perform approximate translation.	Treble CLEF User Communities Workshop (Gonzalo et al. 2009)
The system should alert the user (i) whenever there is more information available in other target languages, (ii) about the quality of MT.	Treble CLEF User Communities Workshop (Gonzalo et al. 2009)
Query translation & refinement	
Include user-assisted query translation facilities.	(He et al. 2003), (Dorr et al. 2004)
Hide user-assisted query translation by default, make it available when things go wrong.	(Petrelli et al. 2003) and most observational studies at iCLEF; Treble CLEF User Communities Workshop (Gonzalo et al. 2009).
Indirect user-assisted query translation that does not involve inspecting foreign-language terms is preferable.	(López-Ostenero et al. 2005a), feedback via suggestion of noun phrases related to the query and taken from the document collection.
Document translation and query translation/formulation/refinement facilities must be consistently designed to fit together.	(López-Ostenero et al. 2005a)
If possible, link structured sources that help mapping the meaning of the query, e.g. named entities, Wikipedia entries.	Treble CLEF User Communities Workshop (Gonzalo et al. 2009)
Personalization	
There must be some support to specify language skills and translation preferences in the user profile.	Treble CLEF User Communities Workshop (Gonzalo et al. 2009)
Image Retrieval	
Combine text-based with content-based facilities.	(Cheng et al., 2005)
Advanced organization features can be appreciated by users.	(Petrelli & Clough, 2006)
Cross-language search should not be offered by default.	(Artiles et al., 2007)
Question Answering	
When user interaction is possible, simple systems may suffice for CL-QA tasks.	(López-Ostenero et al., 2005b); Peinado et al. 2006); López-Ostenero et al., 2008)
Users need more context than in a monolingual setting to assess potential answers.	(He et al., 2005; Figuerola et al., 2005; López-Ostenero et al., 2005b; Navarro et al., 2005)
If feasible, use monolingual IR over a translated document collection as backbone.	(Gonzalo & Oard, 2005; López-Ostenero et al., 2005b; López-Ostenero et al., 2008)

Acknowledgements

Many thanks go to Jacques Savoy for discussions on many of the issues covered in this report. He has also graciously allowed the citation of some highly relevant unpublished work.

References

Abdou, S., Savoy, J. (2006): Statistical and comparative evaluation of various indexing and search models. In AIRS-2006, October 2006, Singapore, Springer-Verlag, Berlin, LNCS 4182, 362-373.

Agirre, E., Di Nunzio, G. M., Ferro, N., Mandl, T., Peters, C. (2008): CLEF 2008: Ad Hoc Track Overview. In Peters, C. (Ed): Working Notes for the CLEF 2008 Workshop. Online at http://www.clef-campaign.org

Artiles, J., Gonzalo, J., López-Ostenero, F., Peinado, V. (2007): Are Users Willing to Sea4rch Cross-Language? An Experiment with the Flickr Image Sharing Repository. Evaluation of Multilingual and Multi-modal Information Retrieval (CLEF2006). LNCS 4730. Springer Verlag.

Ballesteros, L., Croft, W. B. (1997): Phrasal translation and query expansion techniques for cross-language information retrieval. In N. J. Belkin, D., Narasimhalu, P. Willett (Eds.), Proceedings of the 20th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (pp. 84-91).

Bansal, V., Zhang, C., Chai, J. Y., Jin,. R. (2005): MSU at ImageCLEF: Cross-Language and Interactive Image Retrieval. Multilingual Information Access for Text, Speech and Images (CLEF2004). LNCS 3491. Springer Verlag.

Braschler M. (2004a): CLEF 2003 - Overview of Results, CLEF 2003, pages 44-63 (Carol Peters, Julio Gonzalo, Martin Braschler, Michael Kluck (Eds.), Comparative Evaluation of Multilingual Information Access Systems, 4th Workshop of the Cross-Language Evaluation Forum, CLEF 2003, Trondheim, Norway, August 2003. Revised Selected Papers. Lecture Notes in Computer Science, Vol. 3237, Springer, 2004, ISBN 3-540-24017-9)

Braschler, M., Peters, C. (2004b). Cross-Language Evaluation Forum: Objectives, Results, Achievements. In Information Retrieval, Volume 7, Issue 1/2, 7-31, Kluwer Academic Publishers.

Braschler, M. (2004c). Combination Approaches for Multilingual Text Retrieval. In Information Retrieval, Volume 7, Issue 1/2, 183-204, Kluwer Academic Publishers.

Braschler, M., Ripplinger, B. (2004d). How effective is stemming and decompounding for German text retrieval? In Information Retrieval, Volume 7, 291-316, Kluwer Academic Publishers.

Braschler, M., Clough, P. (2008): TrebleCLEF Deliverable 3.1: System Developers Workshop. Available from the TrebleCLEF website www.trebleclef.eu

Buckley, C., Mitra, M., Walz, J., Cardie, C. (1997): Using Clustering and SuperConcepts Within SMART: TREC 6. In TREC 6 Proceedings

Chen, A., Gey, F. C. (2004): Multilingual Information Retrieval Using Machine Translation, Relevance Feedback and Word Decompounding. In Information Retrieval, Volume 7, Issue 1/2, 147-180. Kluwer Academic Publishers

Cheng, P-C., Yeh, J-Y., Ke, H-R., Chien, B-C., Yang, W-P. (2005): Comparison and Combination of Textual and Visual Features for Interactive Cross-Language Image Retrieval. Multilingual Information Access for Text, Speech and Images (CLEF2004). LNCS 3491. Springer Verlag.

Cleverdon, C. (1977) The Cranfield Tests on Index Language Devices. In: K. Sparck-Jones and P. Willett, eds. Readings in Information Retrieval, Morgan Kaufmann, 1997. pp. 47-59.

Clough, P., Müller, H., Sanderson, M. (2005): The CLEF 2004 Cross-Language Image Retrieval Task. Multilingual Information Access for Text, Speech and Images (CLEF2004). LNCS 3491. Springer Verlag. 2005

Clough, P., Al-Maskari, A., Darwish, K. (2007): Provinding Multilingual Accesso to Flickr for Arabic Users. Evaluation of Multilingual and Multi-modal Information Retrieval (CLEF2006). LNCS 4730. Springer Verlag.

Dolamic, L., Fautsch, C., Savoy, J. (2008): UniNE at CLEF 2008: TEL, Persian and Robust IR. In Peters, C. (Ed): Working Notes for the CLEF 2008 Workshop. Online at http://www.clef-campaign.org

Dorr, B. J., He, D., Luo, J., Oard, D. W., Schwartz, R., Wang, J., Zajic, D. (2004): iCLEF 2003 at Maryland: Translation Selection and Document Selection, in Peters et al. (eds.) Comparative Evaluation of Multilingual Information Access Systems. CLEF 2003 Revised Papers, Springer LNCS 3237.

Figuerola, C., Zazo, A., Alonso-Berrocal, J. L., Rodríguez Vázquez, E. (2005): Interactive and Bilingual Question Answering Using Term Suggestion and Passage Retrieval. Multilingual Information Access for Text, Speech and Images (CLEF2004). LNCS 3491. Springer Verlag.

Frakes, W. B. (1992): Stemming Algorithms. In: Frakes, W. B. and Baeza-Yates, R. (Eds.): Information Retrieval, Data Structures & Algorithms, pp. 131-160. Prentice Hall, Eaglewood Cliffs, NJ, USA.

Gey, F. C. (2005): How Similar are Chinese and Japanese for Cross-Language Information Retrieval? Proceedings of NTCIR-5 Workshop Meeting, December 6-9, 2005, Tokyo, Japan

Gonzalo, J., Oard, D. (2005): iCLEF 2004 Track Overview: Pilot Experiments in Interactive Cross-Language Question Answering. Multilingual Information Access for Text, Speech and Images (CLEF2004). LNCS 3491. Springer Verlag.

Gonzalo, J., Peñas, A., Verdejo, F., Peters, C. (2009): Workshop on best practices for the development of Multilingual Information Access systems: the user perspective – The TrebleCLEF user communities workshop report, Del 3.2, TrebleCLEF (EC IST ICT-1-4-1 coordinated action, #215231)

Harman, D (1991): How Effective is Suffixing?. In Journal of the American Society for Information Science, 42(1), pp. 7-15.

He, D., Wang, J., Oard, D. W., Nossal, M. (2003): Comparing User-assisted and Automatic Query Translation, in Peters et al. (eds.) Advances in Cross-Language Information Retrieval. Springer LNCS 2785.

He, D., Wang, J., Luo, J., Oard, D. (2005): Summarization Design for Interactive Cross-Language Question Answering. Multilingual Information Access for Text, Speech and Images (CLEF2004). LNCS 3491. Springer Verlag.

Hull, DA (1996): Stemming Algorithms - A Case Study for Detailed Evaluation. In Journal of the American Society for Information Science 47(1), pp. 70-84.

Karlgren, J., Hansen, P. (2003): SICS at iCLEF 2002: Cross-Language Relevance Assessment and Task Context, in Peters et al. (eds.) Advances in Cross-Language Information Retrieval. Springer LNCS 2785, 2003.

Karlgren, J., Hansen, P. (2004): Continued Experiments on Cross-Language Relevance Assessment, in Peters et al. (eds.) Comparative Evaluation of Multilingual Information Access Systems. CLEF 2003 Revised Papers, Springer LNCS 3237.

Karlgren, J., Gonzalo, J., Clough, J. (2007): iCLEF 2006 Overview: Searching the Flickr WWW Photo-Sharing Repository. Evaluation of Multilingual and Multi-modal Information Retrieval (CLEF2006). LNCS 4730. Springer Verlag.

Llopis, A. F., Vicedo, J. L., Diaz, M. C., Martínez, F. (2003): Universities of Alicante and Jaen at iCLEF, in Peters et al. (eds.) Advances in Cross-Language Information Retrieval. Springer LNCS 2785.

López-Ostenero, F., Gonzalo, J., Verdejo, F. (2005a) Noun phrases as building blocks for Cross-Language Search Assistance. Information Processing and Management, 41 (3), pp. 549-568.

López-Ostenero, F., Gonzalo, J., Peinado, V., Verdejo, F. (2005b): Interactive Cross-Language Question Answering: Searching Passages vs. Searching Documents. Multilingual Information Access for Text, Speech and Images (CLEF2004). LNCS 3491. Springer Verlag.

López-Ostenero, F., Peinado, V., Gonzalo, J., Verdejo, F. (2008): Interactive question answering: Is Cross-Language harder than monolingual searching? Information Processing & Management 44 (1), Special topic issue on User-centered Evaluation of Information Retrieval Systems. pp. 66-81.

Mandl, T., Womser-Hacker, C., Di Nunzio, G. M., Ferro, N. (2008): How robust are multilingual information retrieval systems? SAC 2008: 1132-1136

McNamee, P., Mayfield, J. (2003): JHU/APL Experiments in Tokenization and Non-Word Translation. In Peters, C. (Ed): Working Notes for the CLEF 2003 Workshop. Online at http://www.clef-campaign.org

McNamee, P. (2008): JHU Ad Hoc Experiments at CLEF 2008. In Peters, C. (Ed): Working Notes for the CLEF 2008 Workshop. Online at http://www.clef-campaign.org

Moulinier, I., Williams, K. (2005): Thomson Legal and Regulatory Experiments at CLEF-2005. In Peters, C. (Ed): Working Notes for the CLEF 2005 Workshop. Online at http://www.clef-campaign.org

Navarro, B., Llopis, F., Varó, M. A. (2004): Comparing Syntactic Semantic Patterns and Passages in Interactive Cross Language Information Access, in Peters et al. (eds.) Comparative Evaluation of Multilingual Information Access Systems. CLEF 2003 Revised Papers, Springer LNCS 3237.

Navarro, B., Moreno, L., Vázque, S., Llopis, F., Montoyo, A., Varó, M. A. (2005): Improving Interaction with the User in Cross-Language Question Answering Through Relevant Domains and Syntactic Semantic Patterns. Multilingual Information Access for Text, Speech and Images (CLEF2004). LNCS 3491. Springer Verlag.

Navarro, B. , Moreno, L., Noguera, E., Vázquez, S., Llopis, F., Montoyo, A. (2006): How Much Context Do Yoy Need: An Experiment About the Context Size in Interactive Cross-Language Question Answering. Accesing Multilingual Information Repositories (CLEF2005). LNCS 4022. Springer Verlag.

Oard, D. (2009): "Multilingual Information Access," in Encyclopedia of Library and Information Sciences, 3rd Ed., edited by Marcia J. Bates, Editor, and Mary Niles Maack, Associate Editor, Taylor & Francis.

Oard, D., Gonzalo, J. (2002) The CLEF 2001 Interactive Track. Proceedings of CLEF 2001, Springer LNCS Series, pp. 372-382.

Oard D. W., Resnik P (1999) Support for interactive document selection in cross-language information retrieval. Information Processing and Management, 35(3):363–379.

Oard, D., Gonzalo, J., Sanderson, M., López-Ostenero, F., Wang, J. (2004): Interactive cross-language document selection. Information Retrieval, 7 (1-2), pp. 205-228.

Ogden, W. C., Davis, M. W. (2000): Improving cross-language text retrieval with human interactions. In: Proceedings of the 33rd Hawaii International Conference on System Sciences.

Peinado, V., López-Ostenero, F., Gonzalo, J., Verdejo, F. (2006): UNED at iCLEF 2005: Automatic Highlighting of Potential Answers. In Accesing Multilingual Information Repositories (CLEF2005). LNCS 4022. Springer Verlag.

Petrelli, D., Demetriou, G., Herring, P., Beaulieu, M., Sanderson, M. (2003): Exploring the effect of query translation when searching cross-language, in Peters et al. (eds.) Advances in Cross-Language Information Retrieval. Springer LNCS 2785.

Petrelli, D., Clough, P. (2006): Concept Hierarchy across Languages in Text-Based Image Retrieval: A User Evaluation. Accesing Multilingual Information Repositories (CLEF2005). LNCS 4022. Springer Verlag. 2006

Porter, M. F. (1980): An Algorithm for Suffix Stripping. In Program, 14(3):130-137. Reprint in: Sparck Jones, K. and Willett, P. (Eds.): Readings in Information Retrieval. Morgan Kaufmann Publishers, San Francisco, CA, USA, pp. 313-316

Ripplinger, B. (2000): The Use of NLP Techniques in CLIR. In Peters, C. (Ed): Working Notes for the CLEF 2000 Workshop. Online at http://www.clef-campaign.org

Savoy, J. (2003): Report on CLEF-2003 Multilingual Tracks. In Peters, C. (Ed): Working Notes for the CLEF 2003 Workshop. Online at http://www.clef-campaign.org

Savoy, J. (2004): Combining Multiple Strategies for Effective Monolingual and Cross-language Retrieval. In Information Retrieval, Volume 7, Issue 1/2, 119-146, Kluwer Academic Publishers

Savoy, J. (2005): Comparative Study of Monolingual and Multilingual Search Models for Use with Asian Languages. ACM Transaction on Asian Language Information Processing, 42(2), 163-189.

Savoy, J. (2006): Light Stemming Approaches for the French, Portuguese, German and Hungarian Languages. ACM-SIAC, 1031-1035

Savoy, J. (2009): Unpublished Work. See http://www.trebleclef.eu/ss09_abstract_braschlersavoy.php

Schäuble, P. (1997): Multimedia Information Retrieval. Content-Based Information Retrieval from Large Text and Audio Databases. Kluwer Academic Publishers

Singhal, A., Buckley, C., Mitra, M. (1996): Pivoted Document Length Normalization. In ACM SIGIR Conference on R&D in Information Retrieval, pages 21-29, 1996.

Suzuki, M., Inoue, N., Hashimoto, K. (2001): A method for supporting document selection in Cross-Language information retrieval and its evaluation. Computers and the Humanities, 35(4):421–438

Voorhees, E. (2002): The Philosophy of Information Retrieval Evaluation. In Peters, C., Braschler, M., Gonzalo, J., and Kluck, M. (Eds.): Evaluation of Cross-Language Information Retrieval Systems. Second Workshop of the Cross-Language Evaluation Forum, CLEF 2001, Revised Papers, pp. 355-370.

Zazo, A., Figuerola, C. G., Berrocal, J. L. Fernández, V. (2006): Use of Free On-Line Machine Translation for Interactive Cross-Language Question Answering. Accesing Multilingual Information Repositories (CLEF2005). LNCS 4022. Springer Verlag