Evaluation, Best Practice & Collaboration for Multilingual Information Access

BEST PRACTICES IN LANGUAGE RESOURCES FOR MULTILINGUAL INFORMATION ACCESS

Nicolas Moreau, ELDA, France

Edited by:
TrebleCLEF Consortium

October 2009

www.trebleclef.eu

BEST PRACTICES IN LANGUAGE RESOURCES FOR MULTILINGUAL INFORMATION ACCESS

Nicolas Moreau,
Evaluations and Language resources Distribution Agency (ELDA)
Paris, France

Edited by:
TrebleCLEF Consortium

October 2009

TrebleCLEF: Evaluation, Best Practices & Collaboration for Multilingual Information Access
Coordination Action: 215281 Kickoff: 1 January 2008 Duration: 24 months

Abstract

The first part of this report outlines the various Language Resources and Natural Language Processing Tools needed in order to build a Multilingual Information Access (MLIA) system. The second part provides an in-depth assessment of the current state-of-the-art in this domain. This assessment is based on the results of a series of surveys of communities and individuals active in the field. In the final section, a set of priority requirements are presented which have been established through intensive consultations with system developers, language industry and communication players. A protocol and an action plan are proposed for developing a set of Language Resources and a Toolkit to cover all the technologies needed when implementing systems with MLIA functionality. This minimal set of Language Resources and basic tools/modules is our contribution to the set up of a Basic LAnguage Resource Kit (BLARK) for MLIA.

Preface

The popularity of the Internet and the consequent global availability of networked information sources and digital libraries have led to a strong demand for multilingual access and communication technologies. These technologies should support the timely and cost-effective provision of knowledge-intensive services for all members of linguistically and culturally diverse communities. This is particularly true in the multilingual setting of Europe. Despite recent research advances, there are still very few operational systems available, and these are limited to the most widely used languages.

The Treble-CLEF project was launched to build on and extend the results already achieved by the Cross Language Evaluation Forum (CLEF). The challenge that has been faced is how to best transfer the impressive research results to a wider market place. The aim has been both to support the development and consolidation of expertise in the multidisciplinary research area of multilingual information access (MLIA) and also to disseminate this know-how to the application communities through a set of complementary activities, with the following goals:

- To promote high standards of evaluation in MLIA systems using three approaches: test collections; user evaluation; and log file analysis
- To sustain a MLIA evaluation community by organizing annual evaluation campaigns and providing high quality access to past evaluation results
- To disseminate knowhow, tools, resources and best practice guidelines, enabling system developers to make content and knowledge accessible, usable and exploitable over time, over media and over language boundaries.

In addition to the annual CLEF evaluation campaigns, the project has organised a summer school and number of workshops, and is responsible for the production of a large volume of literature and training material on Research and Development in the MLIA domain. The results are now being transferred to the application communities through the promotion of dissemination events, and via the implementation of a project portal to be made public in Autumn 2009.

Three Best Practice Guidelines will be published:

- Best Practices in Language Resources for Multilingual Information Access
- Best Practices in System and User Oriented Multilingual Information Access
- Best Practices for Test Collection Creation & Evaluation Methodologies

For more information on TrebleCLEF activities and results see http://www.trebleclef.eu/.

Carol Peters, ISTI-CNR, Pisa, Italy
TrebleCLEF Coordinator

Table of Contents

Abbreviations and Acronyms

BLARK	Basic Language Resource Kit
CLIR	Cross Language Information Retrieval
ELDA	Evaluations and Language Resources Distribution Agency
ELRA	European Language Resources Association
ELSNET	European Network of Excellence in Language and Speech
GIR	Geographic Information Retrieval
HLT	Human-Language Technologies
IE	Information Extraction
IR	Information Retrieval
LDC	Linguistic Data Consortium
LR	Language Resource
MLIA	Multi-Lingual Information Access
NERC	Named Entity Recognition and Classification
NLP	Natural Language Processing
POS	Part-Of-Speech
RTE	Recognizing Text Entailment
TC	Text Categorization
WSD	Word Sense Disambiguation

1 Introduction

Language Resources (LRs) are recognized as a central component of the linguistic infrastructure, necessary for the development of Human Language Technologies (HLT), and therefore for industrial development. Other applications may be served by the availability of LRs such as the content industry, cultural heritage safeguarding, etc. The availability of adequate LRs for as many languages as possible and, in particular, of multilingual LRs is a pre-requisite for the development of a truly multilingual Information Society.

The purpose of this report is to describe the state-of-the-art for the Language Resources needed for any kind of Multilingual Information Access (MLIA) system and to assess priority requirements through consultations with system developers, and the language and communications industries. It proposes an action plan for developing a set of Language Resources for all technologies related to MLIA and for the modules employed by those technologies.

By MLIA systems, we mean here any system or component for information access and retrieval in mono- or cross-language mode

Language resources should be interpreted in a wide context: text data, speech data but also modalities and media beyond language such as video, acoustics, images, etc. In this document, language resources also include the basic tools and modules used to process the data.

The document is organized in three main sections.

- The first section is a general presentation of the different elements necessary to build MLIA systems, i.e. the basic tools and language resources, called “MLIA resources” in this document.
- The second section is the result of a survey on MLIA language resources. The goal of this survey is to identify available language resources and specify the sets of LRs that are needed for MLIA system building, according to type, language and media involved.
- The objective of the last section is to identify priority requirements through consultations with language industry and communications players. Finally, taking as a starting point the observed mismatch between existing resources and required priority resources, we propose an action plan for developing new language resources for all technologies related to MLIA and modules employed by those technologies.

2 Main Categories of MLIA Resources

This section gives a general overview of the resources that are needed for MLIA system building. In this document, the term “language resource” covers not only traditional text resources, but also multimedia resources (useful for multimodal IR, combining text with other media) as well as basic natural language processing tools.

The section is organized as follows:

- overview of the MLIA context: definition and main categories of applications,
- overview of MLIA constituents: main categories of modules and tools,
- overview of MLIA resources: main categories of resources.

2.1 MLIA Applications

MLIA deals with all technologies that aim at extracting information from data collections in multiple languages where these languages are not necessarily the same as the language used by the user to formulate the information need (based on keywords, or a natural language query, question, etc...).

The language of the query is called the source language. The language of the document collection is called the target language. In many cases (in particular with web data) the collection itself is multilingual, i.e. contains more than one target language, making the task more challenging.

Below is a non exhaustive list of applications that concern MLIA: All these applications can be developed in a monolingual (source and target languages are the same), bilingual (one single target language, different from the source) or truly multilingual context (several target languages within the same collection or divided over several collections). Main applications are:

- *Information Retrieval* (IR): IR systems allow a user to retrieve the relevant documents which (partially) match his/her information need (expressed as a query) from a data collection. The system yields a list of documents, ranked according to their estimated relevance to the user's query. It is the user's task to look for the information within the relevant documents themselves once they are retrieved.

  An IR system is generally optimized to perform in a specific domain: newspapers, spoken documents, scientific data, patents, law texts, enterprise database, etc. In a multilingual context, we talk of CLIR (Cross-Language Information Retrieval). In the rest of this document we will often use the CLIR and MLIA acronyms indifferently.

- *Question Answering* (QA): Question Answering is a particular form of information retrieval where information needs are expressed as natural language statements or questions. In contrast to classical information retrieval where complete documents are considered relevant to the information need, in Question-Answering specific pieces of information are returned as an answer. Often, automated reasoning is needed to identify correct answers.

  The explosive demand for better information access for a large public of users fosters the need for Research and Development (R&D) for QA systems. The interest of QA is to provide inexperienced users with a flexible access to information allowing them to write a question in natural language and obtain directly a concise answer.

- Other applications concerned by MLIA include: Multilingual Information Extraction, Multilingual Document Filtering, Multilingual Summarization.

In recent years, the impressive growth of available multimedia data (video, photos...) on the web and in professional and private databases has required the development of new information access strategies. New retrieval technologies are required to deal with these data:

- IR in annotated image collections (images with captions, descriptions, etc.).
- IR on multimedia documents combining text and pictures.
- IR in speech transcriptions (information search in transcribed audio-visual documents).
- etc.

Multimedia and audio-visual data are generally processed by multimodal approaches, i.e. IR technologies combining information extracted from text, audio transcriptions, images, video key-frames, etc.

2.2 MLIA Resources

A very general structure of a CLIR system is depicted in Figure 1.

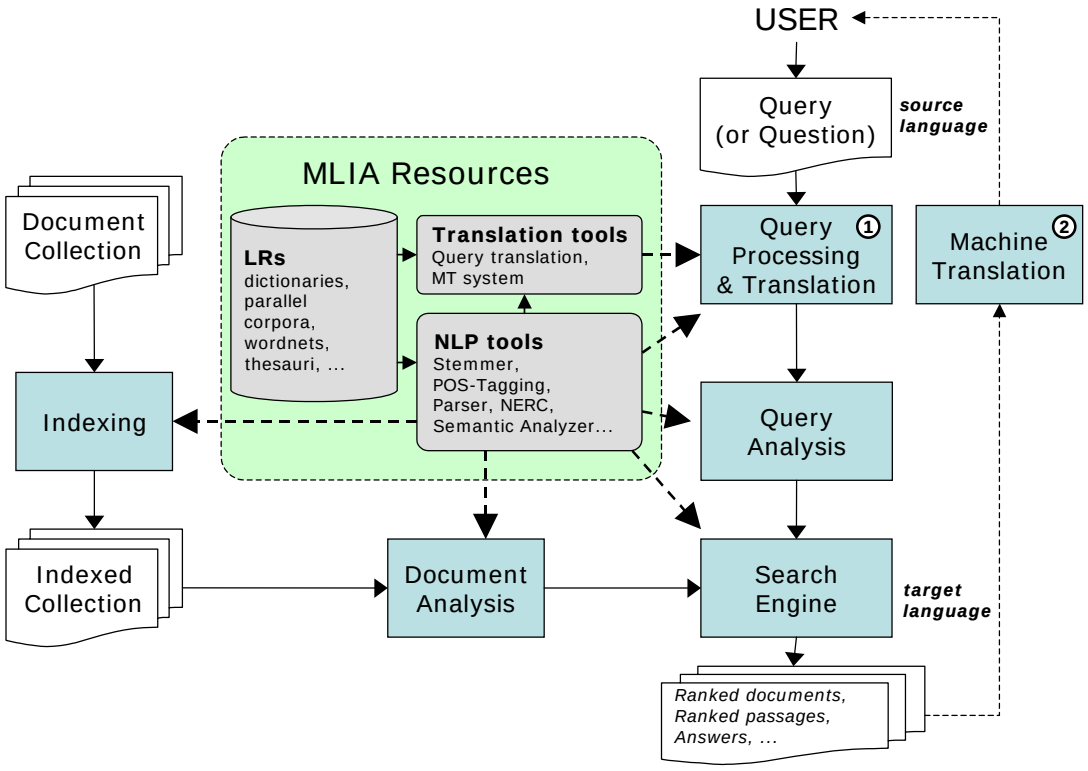

Figure 1: General structure of a CLIR system.

The main elements of a CLIR system are:

- Pre-processing of queries, generally using language dependent techniques.
- Translation of the query (or relevant query terms) into the target language.
- Semantic analysis of the query (extraction of a semantic representation) and comparison with the document representation through the search engine algorithm (computation of a relevance measure).
- Most relevant documents are filtered, selected, ranked and finally displayed to the user (possibly after translation into the source language).
- Before exploiting the system, documents are generally indexed using (language dependent) pre-processing and semantic tools, in order to speed up the IR process.

The design and training of such a system requires two kinds of resources:

- Natural Language Processing (NLP) Tools
- Language Resources

Most of these resources are language specific (others are also domain specific: geographical data, medical data, etc.). The development and upgrading of a multilingual IR system dealing with several languages (and possibly different domains) thus requires many different resources that are expensive to produce, maintain and update.

The following sections describe the main categories of resources and processing tools that are necessary to build a CLIR system. We focus here on the most basic resources and tools needed by core MLIA systems.

2.2.1 Pre-processing

The pre-processing step mainly consists in performing *tokenization* (through a *tokenizer*) and *stop-word removal*. Tokenization is the process of converting a sequence of characters into a sequence of tokens, i.e. deciding where words and sentences start and end. It segments the user input into words and groups words into sentences. This is trivial for humans, but not always for computers (deciding, for example, that a period might end a sentence or not depends on many context factors.)

Stop word removal consists in removing words which are useless for the information search task. This is done using pre-defined and language-specific *stop-word lists*.

2.2.2 Morphological Analysis

Morphological analysis modules determine the base form of words and the morphological information that the inflectional suffixes (or prefixes or infixes) add to the information originating from the base form.

A *morphological analyzer* often consists of a *stemming* module. Stemming is the process of removing any affixes from words, and reducing these words to their roots or to a base form. For example, stemming the English word *computing* produces the base form *comput*. This is the same base produced by the word *computation*. The most famous stemmer algorithm is the well-known Porter's stemmer 2.

Reducing words to their roots or base forms is used for text analysis, indexing and searching.

- Instead of searching for a complete word in a dictionary, only the base form would be searched. This reduces the size of the dictionary.
- Searching for the base form of a word gives a wider search than trying to find an exact match.
- In statistical text analysis, stemming helps in mapping grammatical variations of a word to instances of the same term.

Sometimes, all the tokenization and morphological processing functions are encapsulated in a single *morpho-lexical analyzer* tool.

2.2.3 POS Tagging and Syntactic Analysis

Part-of-speech tagging (POS tagging or grammatical tagging) is the process of assigning grammatical part of speech tags to words based on their context. It is used in particular to tag corpora. A tagged corpus is more useful than an untagged corpus because there is more information there than in the raw text alone. Tags can be used to extract information from the corpus that can then be used for creating *dictionaries* and *grammars* of a language using real language data.

Syntactic analysis (or parsing) is the process of analyzing a sequence of tokens to determine its grammatical structure and major functional relations between words, with respect to a given grammar. Shallow parsing (also chunking, or "light parsing") is an analysis of a sentence which identifies the constituents (noun groups, verbs, verb groups, etc.), but does not specify their internal structure, nor their role in the main sentence. The corresponding tools are called syntactic *parsers*.

2.2.4 Named Entity Recognition

When extracting information from a text, an essential sub-task is to recognize information units like names (including person, organization and location names), and numeric expressions (including time, date, money and percent expressions). Identifying references to these so-called *named entities* in text is called *Named Entity Recognition and Classification* (NERC).

Named Entity Recognition is an important module, required as a tool for almost all the IR or QA system components. A NERC system makes it possible to extract proper nouns as well as temporal and numeric expressions from raw text, tagging each extracted entity with a predefined category (e.g. "Person", "Time", "Organization", "Measure", etc.).

2.2.5 Semantic Parsing

Semantic parsing (also known as semantic role labelling) is the automatic assignment of semantic classes and roles to text, relating syntactic structures to their language-independent meanings.

A *semantic parser* identifies all the predicates in a sentence, and then, identifies and classifies sets of word sequences, that represent the arguments (or semantic roles) of each of these predicates. In simple words, this is the process of assigning a WHO did WHAT to WHOM, WHEN, WHERE, WHY, HOW etc. structure to plain text. This process is necessary in various higher-level NLP tasks (information extraction, question answering, machine translation, etc.), providing them with a layer of semantic structure on top of the syntactic structure.

The training of semantic parsers usually require specific, semantically tagged language resources such as *thesauri*, *semantic networks*, *tree banks*, etc.

2.2.6 Language Resources

Most of the NLP tools are data driven and require particular language resources to be trained, or to process documents and queries within an IR system. The main categories of written language resources are:

- *Monolingual corpora*. A lot of raw, monolingual, un-tagged *corpora* are available. However, a corpus generally has more value for NLP purposes when it is tagged, e.g. POS-tagged corpus or treebank (a parsed corpus, i.e. where sentences have been annotated with their syntactic structures). Annotated corpora are sometimes enhanced with semantic or other linguistic information.
- *Multilingual parallel corpora***.** Large collections of translations of the same texts in different languages are called parallel corpora. Alignments of multilingual parallel corpora at sentence level are prerequisite for the development of many MLIA applications. *Comparable corpora* (i.e. multilingual texts that are not aligned, but address the same domain) are also a useful resource in the MLIA context.
- *Lexicons* (or *dictionaries*). A lexicon gives the vocabulary of a specific language, including its words and expressions. A lexicon organizes the vocabulary of a language according to certain principles (for instance, all verbs of motion may be linked in a lexical network). Some lexicons address a specialized vocabulary (e.g. medicine). In the MLIA context, *multilingual lexicons* constitute a key resource. For instance, a bilingual lexicon or dictionary can be used to translate key words or phrases of a query from one language (source language of the query) to another (target language of the documents).
- *Thesauri* (or *ontologies*). A thesaurus is a database or list of terms (topical search keys). Thesaurus databases are generally arranged hierarchically by themes and topics. In the domain of information technologies, thesauri are sometimes referred to as ontologies. An *ontology* is a formal representation of a set of concepts within a domain. It provides a shared vocabulary, which can be used to model this domain (types and properties of concepts that characterize the domain and the relationships between those concepts).
- *Test collections.* A test collection is a corpus of documents, plus a set of queries that a user might issue to a search engine that has indexed the corpus. In addition, for each query, there is a list of those documents in the corpus that are relevant to the query. Test collections are used by IR researchers to measure the effectiveness of their searching system.

For the training and the evaluation of multimodal IR systems (combining text, images, audio, video, etc.); particular multimedia language resources are required, depending on the context of the application:

- Transcribed speech corpora,
- Annotated image corpora (e.g. pictures with multilingual captions),
- Annotated / transcribed video corpora, etc.

2.2.7 Translation

The primary challenge of a multilingual IR system is to deal with different document and query languages ("crossing of the language barrier").

Translating all documents in the data collection with an MT engine is not always feasible because of the performance of today's MT systems. The most common approach consists in translating the query into the target language (the language of the documents).

In general, a minimal MLIA system will translate the query into the target language before running the search engine (translation module (1) in Figure 1). A complete MLIA system would then translate the retrieved documents, passages or answers into the source language (translation modules (1) + (2) in Figure 1). This additional processing may take the form of machine translation of documents, snippets or document summaries.

Cross-language information retrieval systems in most cases depend upon the availability and the quality of the linguistic resources necessary to translate the query. Depending on the chosen translation approach, these core MLIA resources can be:

- Machine Translation (MT) engines
- Parallel/comparable corpora
- Bilingual dictionaries
- Multilingual thesauri
- Conceptual interlingua

It is to be noted that there are methods (e.g. the *n*-gram matching approach, etc.) that attempt to perform CLIR without using any translation resources.

2.2.8 Language Identification

The design of a truly multilingual system also requires a language identification module to automatically determine the language of a document (or a document passage). This step is necessary in order to apply the corresponding language-specific NLP tools and resources.

Some dedicated *language identifier* tools are freeware, others are commercial tools. This is often included as a functionality of text categorization modules used for indexing the document collections.

2.2.9 Summary

Although it is artificial to make the distinction between NLP tools and language resources (since both are intrinsically linked), the MLIA resources that will be addressed in this document can be described by the following (non-exhaustive) lists of keywords.

NLP Tools

- Morphological analyzers
- Stemmers
- POS taggers
- Parsers
- Named Entity Recognizers
- Semantic Analyzers
- MT systems
- Language Identifiers

Language Resources

- Corpora (mono- & multilingual, tagged, parallel, and multimodal)
- Stopword lists
- Lexicons & Dictionaries
- Grammars
- Ontologies, thesauri, wordnets

3 Survey of MLIA Resources

Another goal of this report is to collect information about the existing MLIA language resources, and assess the needs. This work was done in three phases.

First, a preliminary questionnaire was sent to the participants of the CLEF 2008 evaluation campaign including some questions regarding the future of MLIA R&D and resource developments.

The second phase involved setting up an online survey and contacting MLIA experts and relevant members of the R&D community. This survey aimed at making an inventory of the main MLIA resources and needs.

A list of priority requirements was then defined.

3.1 Preliminary Questionnaire and Final Survey

This section describes the questionnaires that were used and provides some details on the profile of respondents and overall observations.

3.1.1 Questionnaires

A questionnaire was submitted to participants at the end of the CLEF 2008 Workshop in Aarhus. It included some questions regarding the future of MLIA R&D. The respondents gave their opinion on what key MLIA R&D issues are and gave some details on the required resources (corpora and technological modules) to promote these R&D topics in the future. Most of them also mentioned languages or pairs of languages they see as critical for the future development of MLIA systems, and for which more resources are needed:

The second, more detailed questionnaire was specifically set up for this report and publicized as the "TrebleCLEF Online Survey on MLIA Resources".. It was circulated on mailing lists, sent to those working in the MLIA domain, and posted on the Home page of the TrebleCLEF website.

For this second online questionnaire, respondents were asked to give details on the MLIA resources they use to design or run their systems. After describing the applications they are developing or working with, they had to list the main resources and NLP tools used by their systems as well as assess what their needs and priority requirements are in terms of MLIA resources and tools.

The next section gives some details on the profile of respondents to these two questionnaires.

3.1.2 Respondents

The respondents who participated to the questionnaires come from 22 different countries, as shown on Figure 2. They mainly belong to the academic world (78.7%), the others working in independent R&D centres (13.1%) and private companies (8.2%).

The respondents' main centres of interest (as MLIA technology developers or as technology users) in the MLIA domain range from Cross-Language Information Retrieval and Question Answering to Information Extraction, Text Classification and Text Summarization. Most of them are working with both monolingual and multilingual technologies.

Many are focussing on domain specific IR technologies (geographic IR, patent retrieval, etc.) or on multimodal IR (content-based image retrieval, video retrieval, etc.).

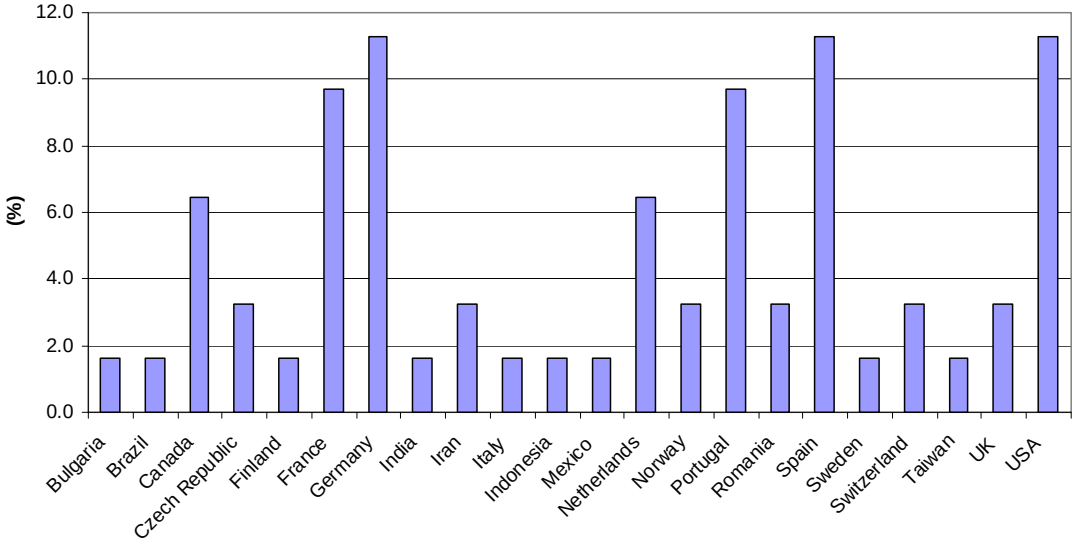

Figure 2: Respondents' countries of origin (%).

3.2 State of the Art

This section provides an inventory of the main MLIA resources and tools identified by respondents. It points to resources and tools without any assessment of their performance or any endorsement of the projects.

Not surprisingly (and especially since the majority of the respondents are from the academic world), most use freely available resources and tools (and make their own resources freely available to the NLP and IR community). In this case, free non-profit and non-commercial use is granted, in general via a GPL license.

This collation of existing data from current MLIA projects was completed by an examination of web and other sources (journals and conference proceedings). In the rest of this section, identified resources are grouped and presented by categories.

This section focuses on the most widely used and well-known resources and tools. A more detailed list of LRs is given in the Appendix where available resources are classified by type and languages covered. This information is synthesized in sections A.2.9 and A.3.7 where 2 matrices are displayed:

- Matrix of languages covered vs. available resources (Table A.2.9),
- Matrix of languages covered vs. available tools (Table A.3.7).

3.2.1 Monolingual Corpora

This section concerns un-annotated and annotated text corpora which constitute by far the most important type of resources, being the foundation for the development of nearly all other resources and tools. Large amounts of text data are needed in all languages for:

- training most NLP tools and IR modules.
- producing other language resources (corpus-based lexicons, grammars, etc.)
- evaluating MLIA technologies (used as test corpora)

There exists an enormous amount of available corpora, freely available or distributed through data catalogues and it is impossible to give here an exhaustive view of existing resources. This section tries to provide an overview of the most popular and easily accessible text corpora.

An important source of good quality and relevant corpora can be the data produced for the different evaluation campaigns dealing with linguistic and information access issues.

In general, these data are made available for free to the project partners or campaign participants. Once the project or campaign is over, some of them are licensed through a commercial catalogue, or made freely available for research purposes. Interested developers should check on the web sites of each project or evaluation campaign of interest to see their distribution policy.

The table below gives an overview of the main IR evaluation campaigns and projects that can provide interesting resources for MLIA purposes.

Among the most commonly used data collections for MLIA are: CLEF test collections for European languages, NTCIR test collections for Asian languages and some TREC test collections for Arabic. FIRE is the main IR initiative for Indian languages.

International IR Evaluation Campaigns & Projects

Name	Description	Org
CLEF	Since it was started in 2000, the Cross-Language Evaluation Forum (CLEF) is a major resource producer for MLIA purposes, as far as European languages are concerned. CLEF datasets of past campaigns include corpora, topics and relevance judgments for many different evaluation tasks, languages (mainly European) and domains (news, web, medical, geographic, annotated images, etc.) CLEF corpora include, for example: • News corpora (used for different evaluation tracks) extracted from newspapers or news agencies: LA Times, Glasgow Herald (English), Frankfurter Rundschau, Der Spiegel, German SDA (German), Le Monde (French), Publico, Folha (Portuguese), etc. • EuroGOV document collection created for WebCLEF (the CLEF web retrieval task). It is a collection of web pages crawled from the European Union portal, EU member state governmental web sites, and Russian government web sites. The corpus contains over 3 million documents written in more than 20 different European languages. • Digital library corpora (for the TEL track): British Library (English), Bibliothèque Nationale de France (French), Austrian National Library (German). • Domain specific corpora: German Target GIRT-4 German English Target GIRT-4 English CSA English Russian Target ISISS Russian • Wikipedia dumps in different European languages. • etc. Availability: Most resources are made available to CLEF participants only, but some of them are planned to be packaged and distributed soon. The CLEF200-2003 evaluation resources are already distributed via the ELDA catalogue: http://catalog.elra.info/. Some of the labs which produced corpora for CLEF make them freely available on the web. Link: http://www.clef-campaign.org/ Languages: mainly European languages, Persian (Farsi).	EC
NTCIR	NTCIR (NII Test Collection for IR Systems) deals with MLIA technologies for Asian languages The NTCIR evaluation tracks include CLIR, CLQA (Cross-Language Question-Answering), The corpora include news articles from different newspapers in Chinese (China Times, Commercial Times, China Times Express, etc.), Japanese (Yomiuri Newspaper), Korean (Korea Economic Daily, Hankookilbo, etc.) and English (Taiwan News, China Times English News, Hong Kong Standard, etc.) Other evaluation corpora consist of collections of patent application documents and web crawls. An overview of the different tasks and the corresponding evaluation corpora is given here: http://research.nii.ac.jp/ntcir/data/data-en.html. Availability: most resources are made available to NTCIR participants only. Link: http://research.nii.ac.jp/ntcir/ Languages: Japanese, Chinese, Korean, English	NII

Name	Description	Org
TREC	Text Retrieval Conference (TREC). TREC mostly deals with monolingual retrieval in English. Until 2001, TREC had an Arabic-English Cross-Language track. Availability: data sets are made available to TREC participants. Most of these resources are distributed by LDC. Link http://trec.nist.gov/ and Languages: Mainly English, Arabic.	NIST
FIRE	Forum for Information Retrieval Evaluation (FIRE). The FIRE website provides test data and other resources (stemmers, bilingual dictionaries, etc.) to the FIRE participants. The evaluations cover several Indian languages (Hindi, Bengali, Marathi, Tamil, Telugu, Punjabi, Malayalam) and English. Availability: data sets are made available to FIRE participants. Some other resources are freely available. Link: http://www.isical.ac.in/~fire/ Languages: Hindi and other Indian languages, English	ISICAL

National IR evaluation programmes generally focus on monolingual research tasks but may also include multilingual resources. In any case, these campaigns also result in the production of valuable resources for the development of MLIA applications. As examples, we mention here two past French evaluation campaigns which addressed information access technologies for the French language.

National Evaluation Programmes

Name	Description	Org
Amaryllis	Launched at the end of 1995, the AMARYLLIS project aimed at evaluating information retrieval software for French text corpora in order to provide a methodology for the evaluation of other similar tools. More specifically, the objective was to create document corpora, questions and answers, in particular for French, similarly to the United States project TREC for English. Text collections in French comprise news articles extracted from "Le Monde"; plus titles and summaries of scientific articles covering every domain. Multilingual text collections comprise documents in 6 languages (French, English, Italian, Spanish, German and Portuguese), extracted from the parallel corpus MLCC which contains documents translated in official European languages. The corpus is divided in two sub-corpora: written questions and debates of the European Parliament. Link: http://www.inist.fr/accueil/profran.htm Languages: French, English, Italian, Spanish, German and Portuguese.	French Government
EQueR	The EQueR (Evaluation campaign for Question-Answering systems) project was part of the Technolangue programme funded by the French Ministry of Research and New Technologies. It provided the resources to carry out a campaign for the evaluation of Question-Answering systems in French. Two text collections were produced: • a corpus consisting of news articles of several years from Le Monde and Le Monde Diplomatique, and press releases and information reports from the French Senate dealing with various subjects. • a domain specific medical corpus consisting of scientific articles and guidelines for good medical practice. Link: http://www.technolangue.net/article.php3?id_article=195 (in French language) Language: French.	French Government

Other evaluation initiatives, less focused on IR issues, have been identified as potential sources of resources, in particular the ones in the following table. All of them have been used by some of the survey respondents for developing MLIA applications.

Other Evaluation Initiatives

Name	Description	Org
MUC	The Message Understanding Conference (IE domain) evaluations used several newswire datasets. Availability: The MUC-6 and MUC-7 datasets are distributed through LDC. Datasets of the previous years are available completely free of charge. Link: http://www-nlpir.nist.gov/related_projects/muc/ Languages: English	NIST
LT4eL Corpora	The Language Technology for eLearning project (LT4eL) created multilingual corpora in the domain of eLearning. Corpora are semantically annotated with concepts, lemmatized, POS tagged. The corpus varies among languages (only partially parallel), the smallest is 200 000 tokens. Availability: free through the LT4eL project. Link: http://www.lt4el.eu Languages: Bulgarian, English, Dutch, German, Czech, Polish, Romanian, Portuguese	LT4eL EC Project
CoNLL	The Conference on Computational Natural Language Learning (CoNLL) created corpora with named entity annotations. Availability: from the CoNLL site. Some data requires a usage license. Some others are free. (e.g.: the CoNLL-2003 data set tagged with NE is freely available at: http://www.cnts.ua.ac.be/conll2003/ner/). Link: http://www.cnts.ua.ac.be/conll/ Languages: Spanish, Dutch, English, German	ACL
TAC	The evaluation workshops of the Text Analysis Conference (TAC) provide large test collections. TAC 2008 has three tracks: Question Answering, Recognizing Textual Entailment (RTE Challenge), and Summarization. Availability: datasets can be obtained through the TAC website (in particular, the corpora of the three first RTE challenges are public.) Link: http://www.nist.gov/tac/ Languages: English	NIST
TIDES	The DARPA TIDES program aims at developing robust technology for trans-lingual information processing (information detection, extraction, summarization and translation). It develops useful resources for MLIA research purposes. Availability: LDC is distributing much of the text, parallel text, lexicons, annotations and other resources used for the TIDES research. Link: http://projects.ldc.upenn.edu/TIDES/ Languages: English, Chinese, Hindi, Arabic, Korean, Spanish...	DARPA

All these projects and evaluation campaigns result in the creation of data collections (as well as associated annotations, topics, relevance judgments, etc.) covering many different languages and specific domains. These are precious resources for the development of MLIA applications.

Some of these data resources are made freely available after the evaluations are over. Others are packaged and distributed via well-known catalogues such as ELDA or LDC. However, some resources are completely available only to the participants in the ad-hoc projects or campaigns. There is therefore a strong need to foster packaging, licensing and dissemination efforts to make these data available to non-participants after the campaigns are over.

As mentioned before, data repositories, such as ELDA or LDC catalogues, are other important sources of appropriate corpora. ELDA and LDC offers a large range of un-annotated and tagged corpora in many different languages. The following table gives an overview of some widely used monolingual corpora distributed by ELDA and LDC.

Resource catalogues

Name	Description	Provider
ELDA Catalogue	Monolingual text corpora distributed by ELDA include (among many others): • The CLEF Test Suite for the CLEF 2000-2003 Campaigns • The AMARYLLIS evaluation package (information retrieval in French text corpora). • The EQueR Evaluation Package (Question-Answering in French) • Italian Treebanks: Italian Syntactic-Semantic Treebank (ISST) / Venice Italian Treebank (VIT) • CELEX Dutch lexical database • The Lancaster Corpus of Mandarin Chinese • TCSTAR corpora (English, Spanish, Chinese) Availability: requires the payment of license fees. ELDA/ELRA Catalogue: http://catalog.elra.info/	ELRA / ELDA
LDC Catalogue	Monolingual text corpora distributed by LDC include (among many others): • Tagged Chinese Gigaword, annotated with full part of speech tags. • Arabic Gigaword • Treebanks (Arabic, Czech, Chinese, Korean...) • Morphologically Annotated Korean Text • Web 1T 5-gram, web data set, contributed by Google Inc., containing English word n-grams (from unigrams to 5-grams) and their observed frequency counts. Availability: requires the payment of license fees. LDC Catalogue: http://www.ldc.upenn.edu/	LDC
BNC	The British National Corpus (BNC) is a 100 million word collection of samples of written and spoken language from a wide range of sources, designed to represent a wide cross-section of British English from the later part of the 20th century, both spoken and written. Availability: license fees. Link: http://www.natcorp.ox.ac.uk/ Languages: Mostly British English	Oxford University

A lot of text corpora are also freely available for research purposes on the web. These resources are generally proposed by the laboratories which produced them with a GPL license. The respondents to the survey mentioned such resources, as shown in the few following examples (many free resources are also listed in the next "Resource Repositories" section).

Free Monolingual Corpora

Name	Description	Provider
Leipzig Corpora	A collection of large corpora 6 in 15 different languages that is free of charge for scientific use (data sources: either newspaper texts or texts randomly collected from the web.) Free for download: http://corpora.informatik.uni-leipzig.de Languages: Catalan, Danish, Dutch, English, Estonian, Finnish, French, German, Italian, Japanese, Korean, Norwegian, Sorbian, Swedish, Turkish	University of Leipzig
OTA	The Oxford Text Archive (OTA) offers many linguistic resources, mainly un-annotated text corpora. Availability: most resources are free. Link: http://ota.ahds.ac.uk/ Language: mostly British English	Oxford University

Name	Description	Provider
Hamshahri and Bijankhan Corpora	The Hamshahri Corpus is a Persian test collection consisting of news texts from the *Hamshahri* newspaper. Right now this collection contains 160,000+ documents. The Bijankhan Corpus is a tagged corpus that is suitable for natural language processing research on the Persian (Farsi) language. This collection is gathered form daily news and common texts. In this collection all documents are categorized into different subjects such as political, cultural and so on. The Bijankhan collection contains about 2.6 million manually tagged words with a tag set that contains 40 Persian POS tags. Both corpora are free for research purposes at: http://ece.ut.ac.ir/dbrg/hamshahri and http://ece.ut.ac.ir/DBRG/Bijankhan/ Languages: Persian (Farsi)	University of Tehran
TwNC	The Twente News Corpus (TwNC) is a collection of text data for language model training. The data includes newspaper data, teletext subtitling and autocues of broadcast news shows and news data downloaded from the WWW (currently more than 300M words in total). Available for research purposes: hltgroup@cs.utwente.nl Languages: Dutch	University of Twente
NEGRA	NEGRA is a syntactically annotated corpus of German newspaper texts. Free for research: http://www.coli.uni-saarland.de/ Language: German	University of Saarland
TGN annotated Wikipedia	A selection of Wikipedia. Articles annotated with the relevant TGN id (TGN is the Thesaurus of Geographic Names). Available from: http://www.doc.ic.ac.uk/~seo01/ (contact: seo01@doc.ic.ac.uk) Languages: English, French, German, Spanish, Portuguese, Chinese, Japanese, Russian, Arabic, etc.	Imperial College London

Finally, many respondents have created data collections themselves, for their own R&D purpose. This is especially necessary when working with less used languages, for which large good quality data sets scarcely exist.

Some of the respondents (working with Swedish, Finnish, Turkish, Indonesian, etc.) have indicated that they built their own data collection, through agreements with local, native speaking newspapers (news corpora); or by using freely available web resources: download of blog data collections, web crawls, Wikipedia dumps, etc. Such corpora cannot be distributed easily but can be created by anyone for internal R&D purposes.

The drawback here is that some types of raw data may be easy to collect, but any further annotation work (to set up a POS-tagged corpus, for instance) requires too much investment and human resources for small academic structures. The participation in collective data production efforts, such as large evaluation campaigns, remains the best way to access large annotated and validated databases.

3.2.2 Parallel Multilingual Corpora

In the MLIA domain, parallel corpora are particularly needed for building bi-lingual vocabulary lexicons and ontologies (in order to develop cross-language query translation and expansion techniques).

A few parallel corpora are freely available for research purposes on the web. But most are distributed via the LDC and ELDA catalogues. Some free and commercial parallel corpora are listed in tables below.

Freely Available Parallel Corpora

Name	Description	Provider
JRC-Acquis	Aligned parallel corpora 7 covering 22 official EU languages. This collection of EU legislative texts changes continuously and currently comprises selected texts written between the 1950s and now. Availability: publicly available on the web. Link: http://langtech.jrc.it/JRC-Acquis.html Languages: 22 EU official languages.	JRC (European Commission)
Europarl	European Parliament Proceedings Parallel Corpus. These are aligned parallel corpora 8 covering official EU languages. This open-source corpus of parallel texts was extracted from the proceedings of the European Parliament, containing up to 28 million words per language. It includes versions in 11 European languages: Romance (French, Italian, Spanish, Portuguese), Germanic (English, Dutch, German, Danish, Swedish), Greek and Finnish. Availability: publicly available on the web. Link: http://www.statmt.org/europarl/ Languages: 11 EU official languages	University of Edinburgh
QALL-ME	The QALL-ME benchmark 10 is a multilingual resource of annotated spoken requests in the tourism domain (Italian, Spanish, English, and German, and their translations into English). Annotations: pragmatic, semantic and QA-based annotations (Expected Answer Type, Expected Answer Quantifier, and Question Topical Target). Further annotations (e.g. named entities, relations) will be made available. Availability: free download (creative common license). Link: http://qallme.fbk.eu/ Languages: Italian, Spanish, English, and German	FBK
MULTEXT-East	Parallel and comparable corpora in Eastern European languages. It includes: - the MULTEXT-East morphosyntactic specifications, lexica, and annotated "1984" corpus - the MULTEXT-East parallel and comparable text and speech corpora; and associated documentation. Availability: partly free. Link: http://nl.ijs.si/ME/ Languages: English, Bulgarian, Croatian, Czech, Estonian, Hungarian, Lithuanian, Romanian, Russian, Slovene, and Serbian.	Jozef Stefan Institute
MuchMore Springer Corpus	The corpus used in the MuchMore project is a parallel corpus of English-German scientific medical abstracts obtained from the Springer web site. The corpus was aligned on the sentence level and consists approximately of 1 million tokens for each language. Abstracts are from 41 medical journals, each of which constitutes a relatively homogeneous medical sub-domain (e.g. Neurology, Radiology, etc.). A tagged version of the corpus is available. Automatic annotations include: POS tags; Morphology (inflection and decomposition); Chunks; Semantic Classes (UMLS, MeSH, EuroWordNet); Semantic Relations from UMLS. Availability: free download. Link: http://muchmore.dfki.de/resources_index.htm Languages: English-German	MuchMore Project (DFKI)
MultiSemCor	MultiSemCor is an English/Italian parallel corpus, aligned at the word level and annotated with PoS, lemma and word sense. Availability: Free for research purposes. Link http://multisemcor.itc.it/ Language: English-Italian	ITC-IRST

Parallel Corpora Available in Catalogues

Name	Description	Provider
LDC parallel corpora	Commercial distribution of parallel corpora via LDC include (among many others): • GALE corpora (Chinese-English and Arabic-English), • ISI Chinese-English Automatically Extracted Parallel Text, • UN parallel text corpora (English-French-Spanish), • Canadian Hansard parallel corpora (English-French), • Bilingual Treebanks: English-Arabic Treebank, English Chinese Translation Treebank. • Hong Kong Parallel Text (English / Chinese): Hong Kong Laws, Hong Kong News, and Hong Kong Hansards corpora. Available at: http://www.ldc.upenn.edu	LDC
ELDA parallel corpora	Parallel corpora distributed by ELDA include (among many others): • EMILLE Lancaster Corpus: (1) monolingual corpora for 7 South Asian languages (Bengali, Gujarati, Hindi, Punjabi, Sinhala, Tamil, Urdu), (2) parallel corpus of text in English with translations in Hindi, Bengali, Punjabi, Gujarati and Urdu. • EMILLE/CIIL Corpus: (1) monolingual corpora for 14 South Asian languages (Assamese, Bengali, Gujarati, Hindi, Kannada, Kashmiri, Malayalam, Marathi, Oriya, Punjabi, Sinhala, Tamil, Telegu and Urdu), (2) parallel corpus of text in English with translations in Hindi, Bengali, Punjabi, Gujarati and Urdu. • JOC MULTEXT corpus (5 European languages), containing raw, grammatically tagged and aligned data from the *Written Questions and Answers* of the Official Journal of the European Community. The corpus contains approximately 1 million words per language: English, French, German, Italian and Spanish. • ARCADE/ROMANSEVAL corpus, containing raw data from the JOC corpus in English and four romance languages: French, Italian, Spanish and Portuguese. • MD Corpus (Le Monde Diplomatique) containing texts in Arabic, Chinese, Greek, Japanese, Persian and Russian manually-aligned with French. A subset for the Arabic-French part was manually annotated with named entities. • CESTA corpora, including several English-French and Arabic-French parallel corpora. • MLCC corpora: (1) monolingual corpora of newspaper articles in 6 European languages (Dutch, English, French, German, Italian and Spanish), (2) multilingual parallel corpus consisting of translated data in 9 European languages (Danish, Dutch, English, French, German, Greek, Italian, Portuguese and Spanish.) • etc. Available at: http://catalog.elra.info	ELDA

The web offers the opportunity to obtain easily free parallel datasets by collecting automatically mined parallel web pages.

3.2.3 Multimodal Corpora

There are still relatively few available corpora of multimodal data (i.e. combining different modalities: text, image, audio, video etc.) that can be used for the evaluation of multimodal IR. Some well-known data sources are mentioned below.

Multimodal Corpora

Name	Description	Provider
Flickr photo collections	The Flickr photo collections are commonly used to create image collections for the development of image retrieval systems. Link: http://www.flickr.com/	Flickr
IAPR TC-12	The image collection of the IAPR TC-12 Benchmark 9 consists of 20,000 still natural images taken from locations around the world (sports, actions, people, animals, cities, landscapes, etc.) Each image is associated with a text caption in up to three different languages (English, German and Spanish). Subsets of IAPR TC-12 were used for ImageCLEFphoto. Availability: free use. Link: http://eureka.vu.edu.au/~grubinger/IAPR/TC12_Benchmark.html Language: English, German and Spanish.	IAPR
CLEF Multimodal Corpora	The CLEF campaigns include some multimodal IR tasks: • Photo retrieval: use mainly the IAPR TC-12 collection (see before). • Medical image retrieval: the data set was made available by the Radiological Society of North America (RSNA) and is also available on the web (see ARRS GoldMiner: http://goldminer.arrs.org). • IRMA collection for medical image classification (11,000 images). • Wikipedia image retrieval: this track uses large collections of Wikipedia images covering diverse topics of interest. • Video retrieval: this pilot track used a collection of 30 hours of dual language videos (Dutch and English) from The Netherlands Institute for Sound and Vision (The Beeld and Geluid Institute) with speech recognition transcripts. Videos are episodes of Dutch television shows, mostly documentaries. Dutch is the main (matrix) language; English is an embedded language which is spoken mainly by interviewees Availability: Most data are available to participants only. Link: http://www.clef-campaign.org/	CLEF
ImageEval Corpora	The ImagEVAL project relates to the evaluation of technologies of image filtering, content-based image retrieval (CBIR) and automatic description of images in large-scale image databases. The main ImageEval benchmark datasets (some including text modalities) are: • Old postcards (~7600 images) • Black & white, colour photographs (~50 000 images) • Dataset for combined text/image retrieval strategies: 700 web pages • Object detection (e.g. Car, tree, …) : 14 000 images • And many others… Availability: Most data are available to participants only. Link: http://www.imageval.org/	Techno-Vision (French program)

Name	Description	Provider
INEX Corpora	The aim of the Initiative for the Evaluation of XML Retrieval (INEX), launched in 2002, is establish an infrastructure and provide means, in the form of a large XML test collection and appropriate evaluation metrics, for the evaluation of content-oriented XML retrieval systems. The resources used for the multimedia track are based on Wikipedia data, in particular: • Wikipedia XML collection: A Wikipedia crawl converted to XML consisting of 659,388 XML documents with image identifiers added to the <image > tags for those images that are part of the Wikipedia image XML collection. • Wikipedia image collection: A subset of 171,900 images referred to in the Wikipedia XML collection is chosen to form the Wikipedia image collection. • Wikipedia image XML collection: This XML collection is specially prepared for the multimedia track. It consists of XML documents containing the images in the Wikipedia image collection and their metadata. Availability: Most data are available to participants only. Link: http://inex.is.informatik.uni-duisburg.de Languages: English, German, French, Dutch, Spanish, Chinese, Arabic, Japanese.	INEX Project
TRECVID Corpora	In 2001 and 2002 the TREC series sponsored a video "track" devoted to research in automatic segmentation, indexing, and content-based retrieval of digital video. Beginning in 2003, this track became an independent evaluation (TRECVID) with a 2-day workshop taking place just before TREC. Availability: Some TRECVID datasets (key frames & transcripts) can be purchased from LDC. Link: http://www-nlpir.nist.gov/projects/trecvid/ Language: English, Mandarin Chinese, Standard Arabic, Dutch.	LDC

Multimodal corpora may also include speech transcriptions, which are of growing interest for the MLIA domain. Mono- or multilingual speech transcription corpora have been produced from within several evaluation projects such as:

- CLEF Cross-Language Speech Retrieval track[1] (CL-SR),
- The NIST Rich Transcription[2] (RT) initiative,
- The TC-STAR project[3] (speech transcriptions of parallel corpora in English, Spanish and Chinese were produced and are distributed by ELDA),
- The AMI[4] and CHIL[5] projects produced collections of videos with transcribed speech (in English).

Most respondents are not satisfied with the existing multimedia resources. There is a strong demand for extending the existing multimodal resources or creating new annotated multimedia and multilingual collections (in particular image collections) to support the growing development of new multimodal IR technologies.

[1] CLEF CL-SR track: http://clef-clsr.umiacs.umd.edu/.

[2] NIST Rich Transcription: http://www.nist.gov/speech/tests/rt/

[3] TC-STAR project: http://www.tc-star.org/

[4] AMI project: http://www.amiproject.org/

[5] CHIL project: http://chil.server.de/

3.2.4 Dictionaries, Lexicons

Multilingual dictionaries are a key resource for the development of MLIA systems. Dictionaries are generally the corner stone of a query translation module.

Some websites provide an updated overview of available dictionaries, such as the Alpha Dictionary Language Directory[6]. This directory selects and adds new dictionaries, grammars, and languages regularly.

Monolingual

Name	Description	Provider
Leipzig Dictionaries	Online search in 48 Corpus-Based Monolingual Dictionaries Availability: free of use. Link: http://corpora.informatik.uni-leipzig.de Languages: 48 languages	University of Leipzig
LDOCE Dictionary	LDOCE is the Longman Dictionary of Contemporary English Availability: commercial. Link: http://www.ldoceonline.com/ Languages: English.	Longman
Monolingual SCIPER dictionaries	Monolingual dictionaries produced in the frame of the EURADIC project. Availability: commercial, distributed via the ELDA catalogue. Link: http://catalog.elra.info Languages: English, German, French, Spanish, Italian	ELDA

Lexicons with semantic information

Name	Description	Provider
HaGenLex HaEnLex	HaGenLex (Hagen German Lexicon) 5 is a domain independent computational lexicon. HaGenLex entries carry detailed morphosyntactic and semantic information. Version for English: HaEnLex (Hagen English Lexicon). Availability: proprietary resources. Link: http://pi7.fernuni-hagen.de/forschung/hagenlex/hagenlex-en.html Languages: German, English	University of Hagen
FrameNet	The FrameNet lexical database contains around 10,000 *lexical units* (a pairing of a word with a meaning), 800 semantic *frames* and over 120,000 example sentences. Availability: commercial or free for scientific use. Link: http://framenet.icsi.berkeley.edu/ Languages: English. FrameNet databases for other languages are available from other sources: - German FrameNet: http://gframenet.gmc.utexas.edu - Japanese FrameNet: http://jfn.st.hc.keio.ac.jp/ - Spanish FrameNet: http://gemini.uab.es/SFN/	University of Berkeley
LT4eL Lexicons	Multilingual lexicons in the domain of eLearning. They cover the concepts in the LT4eL ontology, and includes also phrases Availability: free download. Link: http://www.lt4el.eu Languages: Bulgarian, English, Dutch, German, Czech, Polish, Romanian, Portuguese	LT4eL Project

[6] Alpha Dictionary Language Directory: http://www.alphadictionary.com/langdir.html

Name	Description	Provider
Geonames geographical database	The GeoNames geographical database is available for download free of charge (under a creative commons license). All features are categorized into feature classes. GeoNames is integrating geographical data such as names of places in various languages, elevation, population and others from various sources. Availability: free download (creative commons). Link: http://www.geonames.org/ Languages: many different languages	GeoNames
GoiTaikei	A Japanese dictionary of over 300,000 words most marked using a semantic ontology of 3,000 classes. Availability: commercial (CD for sale online) Link: http://www.kecl.ntt.co.jp/mtg/resources/GoiTaikei Language: Japanese	NTT
MULTEXT Lexicons	Developed in the MULTEXT project, these lexicons are three-column tables: (a) the word-form, (b) the lemma, and (c) the morpho-syntactic information associated to that form (conformant with the MULTEXT/EAGLES specifications.) Link: http://catalog.elra.info Languages: English, French, German, Italian, Spanish.	ELDA

Multilingual Dictionaries

Name	Description	Provider
Bilingual EURADIC dictionaries	Bilingual dictionaries produced in the frame of the EURADIC project. Availability: commercial, distributed via the ELDA catalogue. Link: http://catalog.elra.info Languages: French-English, French-German, French-Spanish, French-Italian, English-German, English-Spanish.	ELDA
DixAF	DixAF is a bilingual French-Arabic dictionary developed at the CNRS. Availability: commercial, distributed via the ELDA catalogue. Link: http://catalog.elra.info Languages: French-Arabic	ELDA
CJK Dictionary Institute	Availability: commercial license. Link: http://www.cjk.org/ Languages: English, Chinese, Japanese, Korean.	CJK
MOT GlobalDix	MOT GlobalDix is a multilingual dictionary in 26 languages. Every indexed word comes with an explanation in English (core language). Availability: commercial. Link: http://www.kielikone.fi/ Languages: 26 languages (with English as core language)	Kielikone Company
Chinese-English Translation Lexicon	The GALE project developed several statistical translation lexicons, including the Chinese-English lexicon. Availability: commercial, distributed via the LDC catalogue. Link: http://www.ldc.upenn.edu Languages: Chinese-English	LDC
CDICT	Chinese-English-Japanese Dictionary. Availability: Free online use. Link: http://cdict.freetcp.com/ Languages: Chinese-English-Japanese	Tamkang University
CC-CEDICT	Freely available Chinese to English dictionary. Availability: Free download (Creative Commons). Link: http://us.mdbg.net/chindict/chindict.php Languages: Chinese-to-English	MDBG

Many respondents have declared that they have purchased multilingual dictionaries but are not always satisfied with these products.

Some laboratories have to develop internally (possibly in the frame of local projects) lexicons that are tuned for particular domain specific uses or less addressed languages. However, this is not regarded as a satisfactory solution, since the development of such new lexicons represents a lot of work. The common and shared development of lexical databases, based on international standards, should be encouraged in the frame of large collaborative research initiatives.

3.2.5 Thesauri, Ontologies, Semantic Networks

Most of the ontologies and thesauri mentioned below exist in many languages, and can thus be used for cross-language retrieval

Name	Description	Provider
WordNet	WordNet is a large lexical database of English. Nouns, verbs, adjectives and adverbs are grouped into sets of cognitive synonyms (synsets), each expressing a distinct concept. Synsets are interlinked by means of conceptual-semantic and lexical relations. This results in a network of meaningfully related words and concepts. Availability: freely and publicly available for download. Link: http://wordnet.princeton.edu/ Languages: English	Princeton University
EuroWordNet	EuroWordNet is a multilingual database with wordnets for several European languages. The wordnets are structured in the same way as the American wordnet for English (Princeton WordNet) in terms of synsets (sets of synonymous words) with basic semantic relations between them. Each wordnet represents a unique language-internal system of lexicalizations. In addition, the wordnets are linked to an Inter-Lingual-Index, based on the Princeton wordnet. Via this index, the languages are interconnected so that it is possible to go from the words in one language to similar words in any other language. Availability: commercial license via ELDA. Link: http://www.illc.uva.nl/EuroWordNet/ Languages: Dutch, Italian, Spanish, German, French, Czech and Estonian (WordNets are currently developed for other languages)	ELDA
Hindi WordNet	Hindi version of WordNet. Availability: commercial license via LDC. Link: http://www.ldc.upenn.edu Language: Hindi	LDC
AWN	The AWN (Asian WordNet) is the result of the collaborative effort in creating an interconnected WordNet for Asian languages (aligned to the English WordNet). Availability: browse online only, download to be available under open source (BSD) license Link: http://asianwordnet.org/ Languages: Thai, Korean, Japanese, Indonesian, Myanmar, Vietnamese, Mongolian, Bengali.	AWN Project
Other WordNets worldwide	The Global WordNet Association provides a list of all available WordNets in the world. Along with the above mentioned ones, WordNets are available, most of the time for free, in many other languages. Link: http://www.globalwordnet.org/ Languages: Arabic, Chinese, Hebrew, Greek, Korean, Persian, Russian, etc.	Global WordNet Association

Name	Description	Provider
MultiWordNet	MultiWordNet is a multilingual lexical database based on Princeton's WordNet. Availability: free for research or commercial licence. Link: http://multiwordnet.fbk.eu Languages: English, Italian (+ possible extensions: Spanish, Portuguese ...)	FBK-irst
MLSN	MLSN is an open source collaborative project to create a Multi-Lingual Semantic Network (which shows synonyms like a thesaurus, but also all the other types of relations between words). Availability: open-source (MIT license). Link: http://dcook.org/mlsn/ Languages: English, German, Japanese, Chinese.	Darren Cook
Roget's Thesaurus	Roget's Thesaurus for English, version 1911. Available at: http://machaut.uchicago.edu/rogets Languages: English	University of Chicago
MeSH	MeSH is the National Library of Medicine's (NLM) controlled vocabulary thesaurus of medical terms. It consists of sets of terms naming descriptors in a hierarchical structure that permits searching at various levels of specificity. Availability: freely available academic research. Link: http://www.nlm.nih.gov/mesh/ Languages: English (+ available translations in German, French, Spanish, Portuguese, Italian, Finnish, and Russian)	National Library of Medicine
UMLS Metathesaurus	The UMLS thesaurus is developed by the National Library of Medicine (NLM). It is a very large, multi-purpose, and multi-lingual vocabulary database that contains information about biomedical and health related concepts, their various names, and the relationships among them. Availability: freely available academic research. Link: http://www.nlm.nih.gov/pubs/factsheets/umlsmeta.html Languages: 17 languages.	National Library of Medicine
LT4eL Ontology	The LT4eL ontology is a knowledge hierarchical repository of concepts in the domain of information technology for end users (i.e. computer science for non-specialists). Based on the OWL Web ontology language standard. Availability: free download Link: http://www.lt4el.eu Languages: Bulgarian, English, Dutch, German, Czech, Polish, Romanian, Portuguese	LT4eL Project
OpenThesaurus	OpenThesaurus is a free German thesaurus and WordNet. Availability: free download Link: http://www.openthesaurus.de Language: German.	OpenThesaurus

Most free resources are in English, and there is a strong need for such resources using less common languages.

Domain specific: geographic locations

Name	Description	Provider
GeoNET	GeoNET Names Server (GNS): The GeoNET geographical ontology: describes place names around the world, in multiple languages. The GNS offers complete files of geographic names for different geopolitical areas (http://earth-info.nga.mil/gns/html/namefiles.htm) Availability: free of use. Link: http://earth-info.nga.mil/gns/html/index.html Languages: Geographical names in many different languages.	NGA
GeoNET-PT	GeoNET-PT is a geographical ontology in Portuguese. Availability: Open source. Link: http://poloxldb.linguateca.pt/index.php?l=geonetpt Languages: Portuguese	Linguateca
Getty TGN	The Getty Thesaurus of Geographic Names (TGN) describes place names around the world. Availability: it requires buying a license Online use: http://www.getty.edu/research/conducting_research/vocabularies/tgn/ Language: mostly in English.	Getty
GeoWordNet	GeoWordNet 11 is a geo-referenced version of WordNet. Availability: free download. Link: http://www.dsic.upv.es/grupos/nle/resources/geo-wn/download.html Language: English	University of Valencia

3.2.6 NLP Tools and Modules

Many powerful and state-of-the-art Natural Language processing (NLP) components are already freely available in the NLP research community: stemmers, lemmatizers, stopword lists, lexicons, chunkers, named entity recognizers, etc. New and improved components are being developed and made available all the time. The components cover the whole range of basic NLP-related processing tasks necessary for the development of IR systems (tokenization, sentence splitting, POS-tagging, lemmatization etc.), but also more complex tasks (parsing, semantic analysis, etc.)

However, most of the components are realized as stand-alone applications without a common and well-defined framework. This is problematic when it comes to integrating components into a processing pipeline for a particular task.

Moreover, linguistic resources can never cover all possible applications. Therefore, adding new languages, new words or adding new information for an existing word is always necessary. A resource is really useful if it is constantly upgraded and adapted to new uses.

Morphological Analysis

Many stopword lists, stemmers and lemmatizers are freely available in the Internet: Snowball stemmers are widely used for monolingual IR and CLIR R&D with European languages (the Snowball stemmer in English is based on the Porter's stemming algorithm 2). About 15 European languages are supported but the coverage of additional languages would be desirable.

Morphological Analyser/Stemmers

Name	Description	Provider
Morphological analyzer	Unsupervised morphological analyzer developed at the Charles University of Prague. Availability: can be obtained for free for academic research by sending an email to: zeman@ufal.mff.cuni.cz Languages: English, Czech, German, Finnish; Turkish, Arabic …	Charles University of Prague
DKPro	The Darmstadt Knowledge Processing Repository (DKPro) consists of a number of scalable and flexible NLP components: tokenizer, stemmer, POS-tagger, etc. Availability: free download. Link: http://www.ukp.tu-darmstadt.de/software/dkpro/ Languages: German	University of Darmstadt
RSLP Stemmer	Portuguese stemmer developed by UFRGS (Universidade Federal do Rio Grande do Sul). Availability: free access. Link: http://www.inf.ufrgs.br/%7Earcoelho/rslp/integrando_rslp.html Languages: Portuguese	UFRGS
JSPELL	Open source morphological analyzer “jspell”. Availability: free for download. Link: http://natura.di.uminho.pt/wiki/doku.php?id=ferramentas:jspell Languages: Portuguese	Universidade do Minho
UNINE Stemmers	Stop-word lists and open source stemming algorithms in many different languages. Availability: open source. Link: http://members.unine.ch/jacques.savoy/clef/index.html Languages: 14 European languages, Arabic, Persian.	University of Neuchatel
Snowball stemmers	Several stemmers implemented using Snowball (a string processing language designed for creating stemming algorithms for IR). It includes the Porter’s stemmer, for English 2. Availability: publicly available. Link: http://snowball.tartarus.org Languages: ~15 European languages (all major European languages)	Tartarus
TextPro	TextPro is a suite of tools oriented towards a number of NLP tasks such as Web page cleaning, tokenization, sentence splitting, morphological analysis, POS-tagging, lemmatization, chunking and named entity recognition. Availability: free for research or commercial license. Link: http://textpro.itc.it/ Languages: English, Italian	FBK-irst
ChaSen	ChaSen is a morphological parser for the Japanese language. This tool for analyzing morphemes was developed at the Matsumoto laboratory, Nara Institute of Science and Technology (NAIST). Availability: open source. Link: http://chasen.naist.jp/hiki/ChaSen/ Language: Japanese	NAIST
Czech morphological analyzer and tagger	A series software tools (Czech morphological analyzer and tagger), developed at the Institute of Formal and Applied Linguistic, Charles University are available with the Prague Dependency Treebank v1.0. Availability: commercial, via the LDC catalogue. Language: Czech.	LDC
Buckwalter Arabic Morphological Analyzer	Buckwalter Arabic Morphological Analyzer. Availability: commercial, via the LDC catalogue. Language: Arabic.	LDC

Syntactic Analyzers / Parsers

Name	Description	Provider
PALAVRAS	Syntactic analyzer PALAVRAS in Portuguese 3 Availability: commercial product. Link: http://beta.visl.sdu.dk/constraint_grammar.html Languages: Portuguese	VISL
XIP	The Xerox Incremental Parsing (XIP) system. Availability: commercial product. Link: http://www.xrce.xerox.com/competencies/content-analysis/robustparsing/ Languages: language independent (XIP grammars have been developed for a number of languages, including French, English and some others are being developed outside Xerox: Japanese, Chinese, German, Portuguese, Czech…)	Xerox
Stanford Parsers	Cf. Stanford NLP toolkits Language: among other languages: English, Chinese, Arabic, German …	Stanford University
Alpino parser	The Alpino parser is an open-source syntactic dependency parser for Dutch. Availability: open source. Link: http://www.let.rug.nl/~vannoord/alp/Alpino/binary/ Language: Dutch	University of Groningen
Machinese Syntax	Machinese Syntax is the syntactic parser of the Machinese tool suite commercialized by the Connexor Company. For sale at: http://www.connexor.eu/ Supported languages: English, French, German, Spanish, Italian, Dutch, Swedish, Danish, Norwegian and Finnish.	Connexor
Charniak Parsers	Statistical parsers developed by Eugene Charniak at the Brown University (Computer Science department). Available for research use at: ftp://ftp.cs.brown.edu/pub/nlparser/ Language: English	Brown University
MINIPAR	MINIPAR is a broad-coverage parser for the English language. Free download: http://www.cs.ualberta.ca/~lindek/minipar.htm Language: English	University of Alberta
PET Parser	The PET parser is a fast HPSG parser. Free for research: http://www.coli.uni-saarland.de/ Language: can use grammars of any language.	University of Saarland
DELPH-IN Repository	DELPH-IN offers an open-source repository of software and linguistic resources including parsers, grammars in many different languages and a toolkit (LKB) for the development of grammars. DELPH-IN also committed itself to a shared format for grammatical representation and to a rigid scheme of evaluation. Link: http://www.delph-in.net/	DELPH-IN Consortium
C&C Parser	The C&C Tools suite include the C&C CCG parser 14. Availability: the email contact for a commercial licence is candc@it.usyd.edu.au. Link: http://svn.ask.it.usyd.edu.au/trac/candc/wiki Language: English.	University of Edinburgh University of Oxford University of Sydney
RADISP	The RADISP syntactic parser 12 is a domain-independent, robust parsing system for English. Availability: distributed freely for non-commercial use Link: http://www.informatics.susx.ac.uk/research/groups/nlp/rasp/	University of Sussex (NLCL group)

NERC

Name	Description	Provider
SIEMES	Named entity recognizer SIEMES Availability: free download. Link: http://poloclup.linguateca.pt/siemes/ Languages: Portuguese	Linguateca
Stanford NER	The Stanford Name Entity Recognizer for English. Availability: free download (GPL license). Link: http://nlp.stanford.edu/software/CRF-NER.shtml Language: English	Stanford University
Yooname	Yooname is a NER system based on and extending the Balie system. Demo: http://www.yooname.com	University of Ottawa
LingPipe NER	NER module of the LingPipe suite. Availability: commercial license, free for academic research. Link: http://alias-i.com/lingpipe Language: English + training data available (usually for research purposes only) in a number of other languages: Arabic, Chinese, Dutch, German, Greek, Hindi, Japanese, Korean, Portuguese and Spanish.	Alias

POS-Tagger

Name	Description	Provider
QTag	QTAG is a probabilistic POS tagger. Academic Free, non-commercial license. Languages: in principle language independent, though it is released only with resource files for English (to use it with other languages a pre-tagged sample text is required for creating the necessary resources. Software to create those resources is included in the distribution.)	University of Birmingham
Stanford POS Tagger	Cf. Stanford NLP toolkits.	Stanford University
TreeTagger	The TreeTagger is a language independent POS-tagger developed within the TC project at the Institute for Computational Linguistics of the University of Stuttgart. Availability: free download. Link: http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger Languages: German, English, French, Italian, Dutch, Spanish, Bulgarian, Russian, Greek, Portuguese, Chinese (easily adaptable to other languages if a lexicon and a manually tagged training corpus are available)	IMS, University of Stuttgart
SVMTool	SVMTool provides an open source POS tagger. Availability: free download (GPL license). Link: http://www.lsi.upc.edu/~nlp/SVMTool/ Language: English, Spanish, Catalan	UPC
TnT POS Taggers	The TnT tagger is a statistical part-of-speech tagger that is trainable on different languages and virtually any tag set. Availability: free for research Link: http://www.coli.uni-saarland.de/~thorsten/tnt/ Languages: German, English (the kit is delivered with language models for German and English; but TnT is not optimized for a particular language. It can easily be adapted to new languages domains)	University of Saarland
ACOPOST	A collection of free open-source POS taggers. Availability: open source. Link: http://acopost.sourceforge.net	SourceForge

Name	Description	Provider
TAIParse	TAIParse is a free standalone part-of-speech tagger and parser. Availability: free download. Link: http://www.textanalysis.com/Apps/POS_Tagger/pos_tagger.html Language: English	VisualText
CLAWS	CLAWS (Constituent Likelihood Automatic Word-tagging System) is a POS tagging software for English text. It has been continuously developed by UCREL (University Centre for Computer Corpus Research on Language) at Lancaster University, since the early 1980s. Availability: licence fee. Link: http://ucrel.lancs.ac.uk/claws/ Language: English	UCREL

Semantic Parsers

The goal of a semantic parser is to assign semantic classes and roles to the constituents of the sentences, in order to provide higher-level NLP modules (machine translation, question classification, etc.) with a layer of semantic structure on top of the syntactic structure.

Semantic parsers are generally trained based on hand-tagged corpora that encode semantic information. Two such corpora are, for instance FrameNet[7] (University of Berkeley) and PropBank[8] (University of Colorado). These corpora make it possible to train supervised machine learning classifiers that can be used to automatically tag vast amount of unseen text with shallow semantic information.

A few free and commercial semantic parsers can be found on the web, including the following ones.

Semantic Parsers

Name	Description	Provider
ASSERT	The ASSERT system (Automatic Statistical SEmantic Role Tagger) is a automatic statistical semantic role tagger, that can annotate naturally occurring text with semantic arguments 17 It was developed by Sameer Pradhan (now working at BBN Technologies). It is based on the PropBank semantics terminology. Availability: freely available for academic research. Link: http://cemantix.org/assert	Sameer Pradhan
Shalmaneser	Shalmaneser (Shallow Semantic Parser) is a supervised learning toolbox for shallow semantic parsing. It is based on the FrameNet semantics terminology 18. Availability: freely available for academic research. Link: http://www.coli.uni-saarland.de/projects/salsa/shal/ Languages: provides end-user pre-trained classifiers for English and German, but the tool is configurable and extendable to other languages.	University of Saarland
Metamap	Along with the UMLS Metathesaurus, the National Library of Medicine (NLM) also offers the Metamap tool to map an arbitrary text to concepts in the UMLS thesaurus (or, equivalently, to discover Metathesaurus concepts in a text). Availability: freely available for academic research. Link: http://mmtx.nlm.nih.gov/	National Library of Medicine
WOCADI	The WOCADI 4 parser (WOrd ClAss based DIsambiguating) is a syntactic-semantic parser (not yet distributed). Demo: http://pi7.fernuni-hagen.de/forschung/wocadi/wocadi_demo.html Languages: German, English	University of Hagen

[7] FrameNet: http://framenet.icsi.berkeley.edu/

[8] PropBank: http://verbs.colorado.edu/~mpalmer/projects/ace.html

Name	Description	Provider
Machinese Semantics	Machinese Semantics is a semantic analyzer that provides semantic role recognition as well as grammatical, lexical and sentential semantic features. Availability: commercial product. Link: http://www.connexor.eu/ Supported languages: English, French, German, Spanish, Italian, Dutch, Swedish, Danish, Norwegian and Finnish.	Connexor

3.2.7 Machine Translation

Some CLIR systems use query translation approaches that do not require a complete machine translation system. Query translation modules are generally based on some of the resources mentioned above: stopword lists and morphological analyzers, source-target bilingual dictionaries, etc.

However, it is possible to use an off-the-shelf (or internally developed) machine translation system (MT system) to cross the language barrier.

A few well-known and widely used free machine translation tools are listed below.

Machine Translation

Name	Description	Provider
EGYPT	The EGYPT Statistical Machine Translation Toolkit was developed at the Center for Language and Speech Processing (CLSP) at Johns-Hopkins University. This is not a ready-to-use MT engine, but a toolkit to develop one's own statistical MT system. In includes GIZA 15, a training program that learns statistical translation models from bilingual corpora. Availability: open source. Link: http://www.fjoch.com/GIZA++.html	Johns Hopkins University (CLSP)
Moses	Moses is a statistical machine translation system 19 that allows to automatically train translation models for any language pair. All you need is a collection of translated texts (parallel corpus). Availability: free (LGPL license). Link: http://www.statmt.org/moses/	Consortium of EU and US labs and institutions
Google language tools	The Google language tools include an online machine translation engine covering many language pairs. Availability: online use. Link: http://www.google.com/language_tools	Google
Babelfish	Babelfish is the online Yahoo! MT engine, based on the Systran technology. It's a rule-based MT engine which makes use of both bilingual dictionaries and linguistic rules empirically designed for specific language pairs. Availability: online use. Link: http://babelfish.yahoo.com/	Yahoo!

Developing MT systems covering a large number of language pairs requires a lot of R&D effort. Many MLIA system developers, including academic developers, purchase commercial MT software. There are a lot of commercial machine translation products on the market, which support many of the most common language pairs.

Among the existing commercial solutions, the survey respondents have mentioned the following ones in particular:

- Systran[9],
- LEC Power Translator[10],
- Promt[11],
- Intertran[12],

Most of these products also have a free of use online version.

3.2.8 Language Identification Tools

Most of the MLIA systems developed by the survey respondents are bilingual (i.e. one single source language vs. one single target language). Only few of them mentioned the use of a language identification tool.

A truly multilingual system, searching information in multilingual collections, clearly needs a language identification module. There are many available solutions, some of them are freely available. We give here a list of some free language identifiers, the most famous being the TextCat Language Guesser[13].

Language Identifiers

Name	Description	Provider
TextCat	The TextCat Language Guesser is a written language identification program based on the text categorization algorithm presented in 16. Availability: free download. Link: http://odur.let.rug.nl/~vannoord/TextCat/index.html Number of languages: 69	University of Groningen
Languid	Languid is a statistical language identifier developed by Maciej Ceglowski. Availability: free download (GPL license). Link: http://languid.cantbedone.org/ Number of languages: 72	Maciej Ceglowski
Mguesser	Mguesser is a standalone part of the libmnogosearch search engine which allows guessing character set and language of a text file. Availability: free download (GPL license). Link: http://www.mnogosearch.org/guesser/ Number of languages: ~50	mnoGoSearch
Lextek Language Identifier	The Lextek Language Identifier supports approximately 260 different languages and character encodings. Availability: commercial product, but a free end-user application is available. Link: http://www.lextek.com/langid/ Number of languages: ~260	Lextek
Balie	The Balie toolkit includes language identification functionalities. Availability: Open source (GPL license). Link: http://balie.sourceforge.net/ Languages: English, French, German, Spanish, Romanian	University of Ottawa

[9] Systran: http://www.systran.co.uk

[10] LEC Power Translator : http://www.lec.com/power-translator-software.asp

[11] Promt: http://www.promt.com

[12] Intertran: http://www.tranexp.com/

[13] TextCat Language Guesser, online demo at: http://odur.let.rug.nl/~vannoord/TextCat/Demo/textcat.html

3.2.9 Resource Repositories

A lot of free resource repositories can be found on the web. Some resource pools are dedicated to one specific language or one specific domain, but most of them offer heterogeneous sets of resources and tools that can be used in developing MLIA applications.

Freely available repositories

Name	Description	Provider
ILPS Resources	The software and resources produced and released by members of the Information and Language Processing Systems (ILPS) group of the University of Amsterdam include: • Arabic blogs dataset: 12,000 Arabic blogs with over 120,300 posts • Blogs and wish lists: a collection of blog and the wish lists of their bloggers • Comment spam in web logs: a toy collection of comment spam in blogs • Concept selection benchmarks for concept-based video retrieval • Historic document retrieval resources: 17th century Dutch: test collections to support historic document retrieval • Information retrieval resources for Bahasa Indonesia: a stemmer, stop word list, as well as two test collections • Source code retrieval • Stemming and stopping in Hungarian • and many others... Availability: free for research. Link: http://ilps.science.uva.nl/resources/	ILPS (University of Amsterdam)
IMS Text Corpora and Lexicon Group	The major focus of the Lexicon and Textcorpora Group at the IMS is the creation of large-scale, high-quality lexicons for natural language applications. The following linguistic resources and tools are offered: • IMSLex, a lexicon for German with morphosyntactic and subcategorization information. • Various kinds of word frequency lists and lexical data with semantic annotations have been compiled, see the Gramotron resources pages. • Tools for automatic text analysis and corpus annotation: for syntactic analysis (automatic annotation with syntactic structures), a range of tools is available, e.g. the stochastic parser LoPar, the LFG grammar, and the YAC system. • Retrieval software for linguistically annotated corpora (e.g. the IMS Corpus Workbench (CQP) and TIGERSearch, a tool for querying syntactically annotated corpora.) • Linguistically annotated text corpora: manually validated 'reference corpora', e.g. a German reference corpus with part-of-speech and lemma information, and a syntactically annotated corpus, the TIGER corpus. • Linguistic Engineering Standards: IMS is involved in various international efforts to standardize linguistic resources and tools: computational lexicons, speech, textual, and multimodal corpora and their annotations, representation formalisms for lexical and syntactic specifications. Availability: free for research. Link: http://www.ims.uni-stuttgart.de/projekte/corplex/	IMS, (University of Stuttgart)

Name	Description	Provider
NLEL Resources	Resources from NLEL (Natural Language Engineering Lab) of the University of Valencia: • Geo-WordNet 0.1: annotation of WordNet 2.0 geographical location synsets with their coordinates • GeoSemCor2.0 : a geographically annotated version of SemCor for WordNet 2.0 • SVM model: An SVM model trained for Arabic Named Entities Recognition on Newswire documents. • ANERCorp: A Corpus of more than 150,000 words annotated for Arabic NER. • ANERGazet: A collection of 3 Gazetteers, (i) Locations: a Gazetteer containing names of continents, countries, cities, etc.; (ii) People: a Gazetteer containing names of people recollected manually from different Arabic websites; and finally (iii) Organizations: containing names of Organizations like companies, football teams, etc. • more than 11,000 Arabic Wikipedia Articles in SGML format (the format adopted in the CLEF and also the one accepted by the JIRS system). • Arabic Wikipedia XML corpus • The CICLing-2002 clustering corpus: This a pre-processed version of 48 scientific abstracts from the CICLing 2002 conference (computational linguistics) which may be used to manually verify the results obtained in the clustering task of narrow domain short texts. • The single-label hep-ex clustering corpus: This corpus is a pre-processed version of the collection of scientific abstracts compiled by the University of Jaén, Spain named hep-ex • The KnCr clustering corpus: This is a new narrow-domain short text corpus in the medicine domain which was constructed by downloading the last sample of documents provided in MEDLINE and selecting only those which are related with the "Cancer" domain. • Blogs clustering corpora: This is a set of corpora made up of discussion lines extracted from two blogs websites: boing-boing and slashdot. • CLiPA corpus: This is a corpus composed by a set of 5 original text fragments (written in English) which have been plagiarised by 9 persons and 5 machine translators (all in Spanish). The corpus has been designed for the development (and test) of Cross-Lingual Plagiarism Analysis applications. Availability: free for research. Link: http://www.dsic.upv.es/grupos/nle/downloads.html	NLEL (University of Valencia)
UPC NLP resources	The NLP research group of UPC offer the following resources: • DiSi - HOPS Flexible Dialogue System for Accessing Web Services • IQMT - Open Source Framework for MT Evaluation. • FreeLing - Open source suite of Language Analyzers. • Omlet & Fries - Open source libraries providing Machine Learning and Feature Extraction facilities. • SVMTool - Open source generator of sequential taggers based on Support Vector Machines. • LangIdent - A Markov-Model based language identifier under GPL license. • Discover de power of MEANING Multilingual Central Repository. • SwiRL - A Semantic Role Labelling system for English (C++, Linux). Bios - A suite of syntactico-semantic analyzers for English. Includes "smart" tokenization, POS tagging (based on TNT), chunking, and Named Entity Recognition and Classification (Java, Linux). • A Machine Translation interface that combines several on-line MT systems to enable translation to/from any available language. Availability: free for research. Link: http://www.lsi.upc.edu/~nlp/web/	UPC

Name	Description	Provider
CPL Resources	The Computational Phonetics & Linguistics (CPL) lab of the University of Saarland provides tools and resources, in particular: TnT (POS tagger), PET (a HPSG parser), NEGRA corpus (a syntactically annotated corpus of German newspaper texts), ... Availability: free for research. Link: http://www.coli.uni-saarland.de/	University of Saarland
NLCL Resources	The NLCL group (Natural Language and Computational Linguistics) of the University of Sussex offers a series of tools (morphological processing, parsers) for English. Availability: distributed freely for non-commercial use. Link: http://www.informatics.susx.ac.uk/research/groups/nlp/resources.php	University of Sussex
SINAI Text Categorization Resources	The SINAI group (University of Jaen, Spain) has made available: • TeCat, a toolkit developed for multi-label text categorization. • HepCorpus, a corpus created for the research of automated multi-label text categorization. It is composed by scientific papers coming from the *High Energy Physics* (HEP) domain, obtained from the CERN Document Server (CDS). Availability: free (GPL license). Link: http://sinai.ujaen.es/wiki/index.php/Recursos	SINAI University of Jaen

Some data and tools repositories are dedicated to a specific language. Below are a few language specific repositories, which are good examples of what could be developed in the future, in particular for less studied languages, for which resource sharing is a necessity (e.g. the ConsILR pool of resources for the Romanian language).

Language Specific Repositories

Name	Description	Provider
Lingateca	Linguateca (distributed language resource centre for Portuguese) is the reference resource repository for Portuguese resources. It includes, among other LRs: • NE annotated corpora of the HAREM project. • WPT 03 and WPT 05 corpora: crawls of the Portuguese Web from 2003 and 2005. • CHAVE collection for IR and QA • CETEMPúblico: corpus of news articles from the Portuguese newspaper "Publico" • CETENFolha : corpus of news articles from the Brazilian newspaper "Folha" • Tokenizer and sentence splitter for Portuguese • etc. Availability: free for research. Link: http://www.linguateca.pt/ Language: Portuguese	Lingateca (Consortium of different Portuguese labs)
CLIA Repository	Cross Language Information Access (CLIA) system for Indian languages. CLIA is a consortium project involving ten Indian universities focusing on IR, Summarization, and machine translation. This is a funded project by Indian government. • Resources developed by various CLIA participants: Wordnet for Hindi, bilingual dictionaries for various pairs of Indian languages, language models, parallel corpora of named entities in various Indian languages… • Tools: Stemmer for Indian languages (Telugu, Hindi), transliteration tool for various Indian languages (including Telugu and Hindi), … Availability: restricted to members of the CLIA consortium. Link: http://search.iiit.ac.in/CLIA2008/ Languages: Indian languages	CLIA (Consortium of 10 Indian labs)

Name	Description	Provider
The ConsILR Pool of Resources	The ConsILR (Consortium for the Romanian Language: Resources & Tools) is a pool of resources dedicated to the Romanian language. It offers to its members some Romanian resources (original or annotated corpora, thesaurus, dictionaries, annotation scheme, etc.) • Romanian Web texts corpora • On-line Romanian-French dictionary • Parallel Romanian-English corpora • Romanian-English dictionary • Corpus of Romanian RST annotated text • MultiSemCor, an English/Italian/Romanian parallel corpus, aligned at the word level and annotated with PoS, lemma and word sense • Sense annotated examples for 39 ambiguous words. • POS Tagging model for the Romanian language • Romanian-English statistical translation system • Lemmatizer for the Romanian Language • NP-chunker for Romanian language texts Availability: require to be a member of ConsILR. Link: http://consilr.info.uaic.ro/ Languages: Romanian	ConsILR (Consortium of different labs)

3.2.10 NLP toolkits

The resource repositories described before generally offer heterogeneous sets of written resources and tools. But there are also NLP tool repositories focusing on the development of more consistent, integrated toolkits. The main ones are described below.

NLP Toolkits

Name	Description	Provider
Linguit Platform	The Linguit Platform offers NLP components for commercial licensing: - Sentence boundary detection, Text tokenization - Morphological processing (stemming, lemmatization) - Grammatical part-of-speech (POS) tagging - Syntactic parser - Semantic analysis component - Named entity recognition and classification (NERC) - Co-reference finder - Toponym resolution (geo-tagger) - Question classification - Indexer: crawls data to construct a searchable index for fast access Availability: commercial license. Link: http://www.linguit.co.uk Language: language independent	Linguit
Machinese	Machinese is a suite of NLP components: tokenizer, POS-tagger, syntactic parser, semantic analyzer, etc. Availability: commercial products. Link: http://www.connexor.eu/ Supported languages: English, French, German, Spanish, Italian, Dutch, Swedish, Danish, Norwegian and Finnish.	Connexor

Name	Description	Provider
Stanford NLP toolkits	The Stanford NLP Group makes several pieces of software available to the public. These are statistical NLP toolkits for major computational linguistics problems. - The Stanford Parser (probabilistic natural language parsers) - The Stanford POS Tagger (maximum-entropy POS taggers) - The Stanford Named Entity Recognizer - The Stanford Chinese Word Segmenter Availability: free download (GPL license). Link: http://nlp.stanford.edu/software/ Language: among other languages: English, Chinese, Arabic, German...	Stanford University
LingPipe	LingPipe is a suite of Java libraries for the linguistic analysis of human language (POS tagging, Phrase chunking, NER). It includes: - Part of Speech Tagging: - Named Entity Recognition: - Document classification Availability: commercial license, free for academic research. Link: http://alias-i.com/lingpipe Language: mostly English	Alias
STILUS	STILUS is a collection of tools for core linguistic processing: parsers, POS taggers, syntactic parsing, disambiguation, text classification, Named Entity extraction... Availability: commercial product. Link (only in Spanish): http://www.daedalus.es/productos/stilus/ Languages: Spanish (Europe, South America, United States), Catalan, Basque, Galician, English, French, Portuguese, German, Italian, Polish, Bulgarian, Arab, Russian	DAEDALUS
FreeLing Package	FreeLing is an open source suite of NLP tools. The FreeLing package consists of a library providing language analysis services: Text tokenization, Sentence splitting, Morphological analysis, NERC, PoS tagging, WordNet based sense annotation... Availability: free download (GPL license). Link: http://garraf.epsevg.upc.es/freeling/ Languages: Spanish Catalan, Galician, Italian, and English.	UPC
NLTK	NLTK — the Natural Language Toolkit — is a suite of open source Python modules, linguistic data and documentation for research and development in natural language processing, supporting dozens of NLP tasks NLTK includes a large variety of NLP tools: tokenizers, stemmers, taggers, parsers, semantic interpretation modules, WordNet, etc. The NLTK website also offers written resources at http://www.nltk.org/data (it includes around 60 resources: corpora, toy grammars, trained models, etc.) Availability: free (creative commons). Link: http://www.nltk.org/	The NLTK consortium
Balie	The Baseline information extraction (Balie) is a multilingual IE toolkit including the following modules: • language identification • tokenization • sentence boundary detection • named-entity recognition Availability: Open source (GPL license). Link: http://balie.sourceforge.net/ Languages: English, French, German, Spanish, Romanian	University of Ottawa

Name	Description	Provider
C&C Tools	The C&C Tools suite 13 consists of the C&C CCG parser (including the computational semantics tool, Boxer) and the C&C taggers (the tools also use the morphological analyser *morpha*). These tools have been developed by James Curran and Stephen Clark (C&C). Boxer generates semantic representations. It has been developed by Johan Bos and is distributed with the C&C tools. The suite also includes *Nutcracker*, a system for recognising textual entailment, developed by Johan Bos. Availability: free licence for non commercial purposes (contact for a commercial licence: candc@it.usyd.edu.au). Link: http://svn.ask.it.usyd.edu.au/trac/candc/wiki Language: English.	University of Edinburgh University of Oxford and University of Sydney
CCG Toolkits	The CCG group (Cognitive Computation Group) of the University of Illinois offer different NLP toolkits of interest, including POS taggers, Named Entity taggers and co-reference resolvers. Link: http://l2r.cs.uiuc.edu/~cogcomp/software.php	CCG (University of Illinois)
OpenNLP	OpenNLP is an organizational centre for open source projects related to natural language processing. OpenNLP hosts a variety of java-based NLP tools which perform sentence detection, tokenization, pos-tagging, chunking and parsing, named-entity detection, and coreference. The website also provides a series of web links to other NLP toolkit projects. Availability: free download. Link: http://opennlp.sourceforge.net/ Languages: English, German, Spanish, Thai.	SourceForge

3.2.11 IR Programming Tools

Although this document does not deal with the development of MLIA systems per se, we include here a section on the IR programming tools and packages. These packages can be seen as an important part of the set of "basic resources" needed to build a MLIA system, and they mostly also include basic NLP tools.

We give below a few examples of open-source toolkits, derived from our survey. They are used to develop CLIR engines and multilingual QA systems. The most popular tools are by far the Lucene and Lemur toolkits. We also mention here a few image retrieval toolkits (in particular FIRE) that can be used to design cross-language multimodal image retrieval systems.

IR Toolkits

Name	Description	Provider
Lucene	Apache Lucene is an open source text search engine library written entirely in Java. Availability: open source. Link: http://lucene.apache.org	Apache Software Foundation
Lemur	The Lemur Toolkit is an open-source toolkit designed to build IR systems. Availability: open source. Link: http://www.lemurproject.org/	UMass and CMU
Terrier	Terrier (Terabyte Retriever) provides a platform for the development of IR applications. Availability: open source. Link: http://ir.dcs.gla.ac.uk/terrier/	University of Glasgow

Name	Description	Provider
GATE	GATE is an open source toolkit for IE and text mining. Availability: open source. Link: http://gate.ac.uk/ Languages: English, French, German, Spanish, Italian, Chinese, Arabic, Romanian.	University of Sheffield
Cheshire II	The Cheshire II is an Information Retrieval toolkit allowing to design mono- or cross-language IR systems. Availability: open source. Link: http://cheshire.berkeley.edu Languages: English, French, German, Portuguese, Spanish, Russian	University of Berkeley
CWB	The IMS Open Corpus Workbench (CWB) is a collection of tools for managing and querying large text corpora (100 M words and more) with linguistic annotations. Availability: open source. Link: http://cwb.sourceforge.net	IMS Stuttgart
JWPL	JWPL (Java Wikipedia Library) is a free, Java-based application programming interface that allows to access all information contained in Wikipedia. Availability: open source. Link: http://www.ukp.tu-darmstadt.de/software/jwpl/ Note: a version for Wiktionary, named JWKTL (Java Wiktionary Library), will also be available soon (http://www.ukp.tu-darmstadt.de/software/jwktl/).	University of Darmstadt
FIRE	The Flexible Image Retrieval Engine (FIRE) is an image retrieval system. Availability: open source (GPL license). Link: http://www-i6.informatik.rwth-aachen.de/~deselaers/fire/	RWTH Aachen University
Caliph & Emir	Caliph & Emir are MPEG-7 based Java prototypes for digital photo and image annotation and retrieval supporting graph like annotation for semantic metadata and content based image retrieval using MPEG-7 descriptors. Availability: open source. Link: http://www.semanticmetadata.net	SemanticData

4 Priority Requirements for MLIA Resources

In order to ensure new advances and real breakthroughs in the domain of multilingual information retrieval it is necessary to constantly improve the existing NLP tools, to enlarge and enrich the textual and multimodal data resources used by data-driven components, and to develop new paradigms. One of the goals of this report is to disseminate the information we collected about present and future needs regarding MLIA language resources. This section provides a list of the main needs identified, derived from the survey and from information collected on the web. In Section 4.2 an action plan is proposed to fulfil those needs.

4.1 Resource Needs

Multimodal CLIR has become a major research issue in the MLIA field. There is a strong and growing interest in exploiting and combining different modalities (text, image, etc.) to retrieve images, audio-visual documents and more generally multimedia documents in a multilingual context. In particular, access to pictures mixed with text (think of the large amounts of multilingual materials combining text with photographic material and/or with illustrations) was pointed out as one major key R&D issue, in addition to multi-lingual and multimodal topic detection and tracking in news and media.

Another major issue is how to transfer research results and prototypes into practical MLIA applications fitting market needs and requirements. This challenge has received more and more attention in the past few years, with digital content search initiatives launched by US companies (Google, Yahoo, MSN), followed by similar initiatives in Europe (the European Commission launched an effort aimed at building The European Library, the French government has just started the Quaero project to develop a multimedia search engine, THESEUS is another research program initiated by the German Federal Ministry of Economy and Technology (BMWi), with the goal of developing a new Internet-based infrastructure in order to better use and utilize the knowledge available on the Internet, etc.) and in Asia. A crucial issue is to "compare" the achievements of the lab prototypes and the commercial products and to address the issue of transferring the lab prototypes to the real world, in particular from the resources perspective.

Many respondents to our web based survey insisted on the fact that new resources are needed to cover domain-specific MLIA technologies, which may then better comply with users' needs and correspond to real application scenarios. Such specific resources may concern, for instance:

- Multimodal MLIA, as already mentioned, e.g. Content-based image retrieval combined with multi-lingual text, CLIR on "noisy data", such as automatic speech transcriptions, OCRized texts, etc.
- MLIA on web data (machine translation of web pages, development of web-specific QA systems): such web specific CLIR systems would be particularly useful for users that are interested in documents written not only in their native language, but also in languages they are able to understand.

The survey respondents also mentioned other domain-specific MLIA fields, such as:

- Cross-language patent retrieval,
- MLIA for enterprise databases and Intranets (global companies require multilingual IR to manage their own data and archives),
- MLIA for "intelligence data" (e.g., multilingual search in web sites and forums, on audio data recordings (in particular telephone conversations), etc.),
- MLIA for personal information management: many people need to manage documents written not only in their own native language, but in languages they are able to understand,
- MLIA on encyclopedia material,
- MLIA in the domain of E-learning,

4.1.1 Key Languages

Respondents mentioned the importance of focusing on the needs in terms of particular languages or language pairs. Our survey asked which languages were considered as critical for the future development of MLIA systems. Languages which have been cited as the most important are shown on Figure 3.

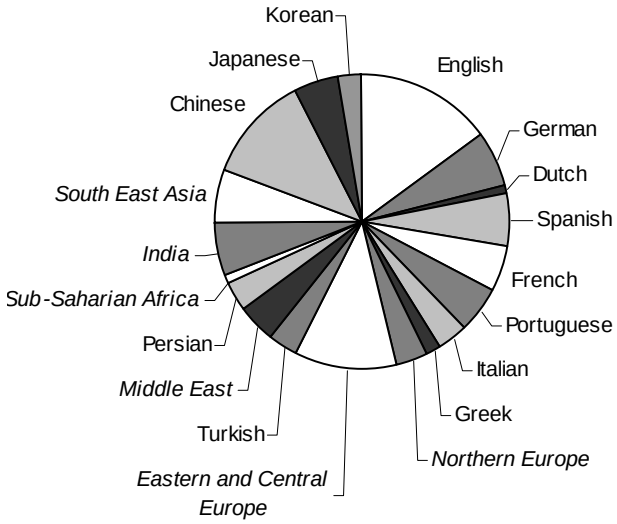

Figure 3: Most needed languages (%).

It is hard to derive clear conclusions from these figures because they are biased by the particular research interests of the respondents, who mainly work with European languages (the fact that English is the most cited language is due to its use as the core language in most MLIA research and evaluation efforts). However, the following observations can be made:

- There is a demand for resources covering the transition from major Western European languages to Eastern and Central European languages, and vice versa. Along with Russian, a lot of less resourced languages were mentioned among the languages from the new EC countries and from the Balkan region. It was pointed out that a prerequisite for MLIA deployments is the development of monolingual Language resources, in particular large monolingual document collections that are missing for most of these languages.
- Of course new resources are also needed for the major Western European languages. From the survey, this is particularly the case for Portuguese which seems to be still underrepresented in terms of linguistic resources, although Portuguese can be seen as a key language since it is one of the languages with the largest number of native speakers. Such assumption could be also stated for other languages and language genres. More semantic resources, especially word-net like ones, were particularly mentioned.
- As far as interactions between European languages and Asian and Middle East languages are concerned, it is no surprise to have English as the priority language when developing such MLIA language directions. Mandarin Chinese is of course the most needed Asian language in terms of language resources.
- Outside the most spoken languages of Europe (English, French, German, Italian, Spanish) and Asia (Chinese and Japanese) or the additional official languages of the United Nations (Arabic and Russian), resources in terms of parallel corpora are still very difficult to obtain (so are the corresponding technologies).

In particular, the languages of the Indian subcontinent have received very little attention from the international community until recently (although 6 of the top 25 languages by numbers of speakers are from the Indian Subcontinent: Hindi, Bengali, Urdu, Telugu, Punjabi, Tamil). Some new Indian initiatives such as CLIA and FIRE[14] are now trying to fill the gap, but virtually nothing has been done in the Western labs with these languages, with the exception of the CLEF2007 experiments with languages from the Indian subcontinent[15].

Interest is also slowly growing in the MLIA field for Persian (Farsi), as shown by the CLEF2008 task for this language [16] and South Eastern languages (Indonesian, Thai, Vietnamese, etc.).

In the Appendix we have classified the available resources and tools according to the covered languages. This information is synthesized in the tables of sections A.2.9 and A.3.7 where the following information is displayed:

- Matrix of languages covered vs. available resources (Table A.2.9),
- Matrix of languages covered vs. available tools (Table A.3.7).

This makes it possible to identify the most important needs in terms of language resources. It is obvious in particular, that large linguistic areas in Asia (especially emerging powers like Persia, India or China) are not very well covered compared to their huge number of speakers. More effort should be put on resource development and/or on resource distribution for these languages (in some cases resources exist but are not made available to the world wide community).

4.1.2 Key Resource Needs

Some information can be derived from the survey and the inventory regarding the resources required (corpora and basic tools) to promote R&D for the above-mentioned key languages, and key research topics. The main issues are:

More domain specific resources (multilingual corpora, domain specific thesauri) addressing real application scenarios, in particular: more multilingual web resources and multilingual / multimodal data collections.

- Large parallel and aligned corpora for more languages (not only English vs. another language) and more domain-specific ones, in order to build translation resources (MT engines, multilingual dictionaries, multilingual thesauri, etc.)
- Development of better WordNets for some languages (in particular, Portuguese and German have been mentioned as requiring more resources of this type).
- Easier access to the valuable resources produced in the framework of evaluation campaigns (in particular training and test corpora, queries, etc.).

Domain specific resources

Corpora coverage of additional genres is an important requirement. MLIA research to date has exploited mostly news corpora in textual form – little has been done with web pages or scientific and technical corpora (i.e. patents, case law).

Past experiments have shown that performance of CLIR methods depends on the availability of high-quality translation resources for particular domains. Severe performance degradation has been observed when using a general-purpose training corpus or a generic commercial machine translation system, versus a domain-specific training corpus 20. For instance, CLIR systems' proven ability to rank news stories does not transfer readily to other genres such as medical journal articles. When

[14] www.isical.ac.in/~clia/

[15] http://www.clef-campaign.org/2007/working_notes/CLEF2007WN-Contents.html

[16] http://www.clef-campaign.org/2008/working_notes/CLEF2008WN-Contents.html

moving away from news into more technical domains, the most common translation of a particular term may no longer be the appropriate one.

Therefore, the effective performance of MLIA systems generally depends on the availability of high-quality "translation resources" for particular domains. In this context, many new resources need to be developed to cover the domains of interest in terms of realistic application scenarios, including:

- Multimodal data collections (see below),
- Multilingual web collections to tune and evaluate web MLIA systems.
- Corpora covering new languages and geographical areas, as well as new gazetteers (*geographical dictionaries, giving information about places and place names*) are required to further develop GIR systems (Geographic Information Retrieval),
- Multilingual patent corpora for the development of patent search applications,
- Plus corpora addressing any specific domain that may correspond to an identified real user need.

Along with multimodal resources which are addressed below, the development of multilingual web resources is considered a priority. A lot of work is clearly needed to define suitable criteria for the construction of more R&D Web corpora which are really multilingual. It is a difficult challenge because of the low percentage representation beyond the top ten languages, as well as of the lack of standards for character and font representation for many other languages. Chinese has at least two major representations (GB and BIG5) and Japanese three, while for Indian subcontinent languages standards are only beginning to be developed (i.e. each site has its own font and internal character representation).

Some work has been done in CLEF (WebCLEF tracks) and in the NTCIR Web IR tracks, but the corpora developed cover a limited set of languages and domains (e.g. the EuroGOV document collection created for WebCLEF is a collection of web pages crawled from the European Union portal, and some European state governmental web sites).

Multimodal resources

Nowadays, classical textual IR techniques cannot cover all the needs in terms of information access. The market is moving towards multimodal information access to cope with the growing amount of multimodal data (image, audio, and video) available on the Web and in private databases. In the perspective of building universal search engines, multilinguality is also of increasing interest in the multimedia IR domain. Obtaining large new multimodal and multilingual data collections for which widespread evaluation benchmarking is a practical and important step that needs to be addressed.

In text retrieval, it has been relatively straightforward to obtain large collections of newspaper texts because the copyright owners have seen their own interest in the development of MLIR systems and it has been easy to convince them to let us reuse the raw text that has been published years before (they are basically vendors of information and not of texts). However, image, video, and speech collection owners do see great value in their collections (which hardly loose their value over time and become out-dated) and consequently are much more cautious in releasing their content.

Useful new video collections could include news video (in multiple languages), collections of personal videos, and possibly movie collections with multilingual speech tracks and subtitles.

Image collections would include image databases (maybe on specific topics) along with annotated text. The use of library image collections should also be explored. One critical point here is to avoid building artificial collections on which evaluated systems would perform well, while they would not necessarily work with more general image collections.

Such collections are difficult and costly to produce. The challenge is to make them easily available (publicly available or distributed through a catalogue at a reasonable price) to the MLIA community.

One obvious solution is to further encourage collaborative research and evaluation efforts combining the multilingual and multimodal issues. Projects producing image and video databases with

appropriate relevance judgments could make more and more resources available to external groups for research purposes. Several old and new joint research initiatives should contribute to providing valuable resources to the MLIA community (CLEF multimodal tasks, TRECVID, Quaero, Chorus, etc.).

Another relevant issue related to multimodality is cross-language information retrieval in "noisy" textual data such as automatic transcriptions of spoken documents or OCRized texts (such transcriptions still have some errors even when they are manually checked). In particular, most speech retrieval systems are tuned to a single target language. Thus, multilinguality is a topic that still needs to be addressed with respect to the creation of ad hoc multilingual data collections.

Parallel corpora

MLIA capacities for many specific language pairs are limited by the lack of translation resources. First of all, large parallel and aligned corpora are needed for these language pairs in order to build the corresponding translation resources (MT engines, multilingual dictionaries, multilingual thesauri, etc.) given the current trend of corpus based (and data driven) approaches for all these techniques.

Parallel corpora are needed to cross the language barrier between well-covered European and Asian languages, without using English as an intermediate language (i.e. allowing direct training of German-Chinese or Spanish-German query translations for instance).

Parallel corpora are also needed to align English with less-resourced languages (like Persian, Hindi, South Eastern Asian languages, etc.).

To allow the construction of bilingual transfer dictionaries and thesauri based on these parallel corpora, tools are needed for extracting terminology and for automatic construction of the semantic relations. If bilingual text corpora are available in a domain, tools for computer aided building of transfer dictionaries could be developed.

Semantic resources

New IR applications include more and more (basic) semantics: automatic classification, Named-Entity detection, geolocation, opinion analysis, etc.

Databases of lexical semantic relations as general as possible are needed in a many languages for monolingual reformulation using classical relations like synonyms, narrower terms, broader terms and also more precise relations like *part of*, *kind of*, *actor of the action*, *instrument of the action*, etc., such as is being created for English in WordNet.

Named-entity tagged data collections, as well as new WordNets (or improved versions of existing ones) are needed for most languages. More effort should also be put on semantic parsers and semantic analysis modules.

4.2 Action Plan

This section proposes an action plan to implement the identified and needed resources within the next few years. It consists in the following main steps.

1) Conduct BLARK (Basic LAnguage Resource Kit) studies for a selected set of key technologies and languages.
2) Identify real user needs through specific surveys that help to draw conclusions on future key applications and the required resources (e.g. domain specific resources and tools).
3) Ensure a large consensus on the resources needed and derive an agenda to make them available. Although BLARK is seen as the minimal set of language resources that is necessary to do precompetitive research, one could extend it to prototypes and commercial products (what ELDA introduced as ELRAK or the Extended Language Resource Kit).

4) Elaborate a generic strategy that can be applied to any specific languages.

5) Foster cross-disciplinary collaborative creation efforts for these resources (parallel corpora, WordNets, etc.) for (all) languages and key domain specific issues.

6) Elaborate efficient distribution strategies for the resources built: production of evaluation packages, clearing of IPR issues, involvements of resource repositories, etc.

Although many media and languages can be lucrative enough to be tackled by commercial organizations, there is no guarantee that they would do that in a framework that would safeguard interests of all technology players and it is clear that they would not pay attention to less lucrative and minority languages.

4.2.1 BLARK Matrices for MLIA

The BLARK concept (Basic LAnguage Resource Kit) was defined in a joint initiative between ELSNET (European Network of Excellence in Language and Speech) and ELRA (European Language Resources Association) 21 and first launched with a Dutch initiative called Dutch Human Language Technologies Platform that was begun in April 1999. Then, in the framework of the ENABLER thematic network (European National Activities for Basic Language Resources -Action Line: IST-2000-3.5.1), ELDA elaborated a report 22 defining a (minimal) set of LRs to be made available for as many languages as possible and mapping the actual gaps that should be filled in so as to meet the needs of the HLT field. A good illustration of the BLARK concept is the work done to gather information about Dutch/Flemish language technology modules and data 23, or about Arabic resources within the NEMLAR project 24.

In particular, ELDA has been working at implementing BLARK matrices[17], to highlight the gaps with regards to LRs needed for specific applications and for as many languages as possible. In order to understand the needs in a clearer and more complete way, ELDA has decomposed the BLARK concepts into two matrices with a list of potential applications and a list of potential modules to be cross-linked with the LRs needed (considering specificities of the corresponding languages if any). The approach consists in identifying the specific needs for a given language by defining the two matrices:

- An 'Applications/Modules' table showing the dependency of each application on the various modules and the level of these dependencies.
- A 'Resources/Modules' table showing the level of need for language resources by each specific module.

In both cases each need (i.e. each cell in the matrices) has to be assessed as important (+), very important (++), essential (+++) or irrelevant (void cell).

One of the goals of this study was to work on an instantiation of a BLARK for MLIA applicable in principle to all languages. The generic BLARK concept, originally defined for HLT in general, was adapted to the specific needs of the MLIA field. The following tables (Table 1 and Table 2) represent a first tentative to instantiate MLIA specific BLARK matrices.

The first matrix (Table 1) shows the correspondence between the main types of MLIA applications and constituents and the necessary modules for building those applications and constituents (as identified in the previous sections).

[17] BLARK website: http://www.blark.org/

Abbreviations:

CL-IE *Cross-Lingual Information Extraction*

CL-IR *Cross-Lingual Information Retrieval*

CL-QA *Cross-Lingual Question-Answering*

Applications / Modules	CL-IE	CL-IR	CL-QA
Sentence Boundary Detection	+	+	+
Tokenizer	++	++	+++
Morphological Analyzer (deriv., stemm., diacritic, …)	++	++	+++
POS Tagger	+++	+++	+++
Chunker (Shallow Parser)	++	++	++
Named Entity Recognizer	+++	++	++
Word Sense Disambiguation	++	++	++
Syntactic Analyzer	++	++	+++
Semantic Analyzer (incl. coreference resolution)	+++	++	+++
Language Identifier	++	++	++
Translation (MT, query translation…)	+++	+++	+++

Table 1: Application vs. Modules

The second matrix (Table 2) shows the language resources that are necessary in order to build the afore-mentioned MLIA modules. In order to make the correspondence clear we are using the same list of modules in the left hand side of the two tables.

Resources / Modules	Stop word list	Un-annotated Corpora	Annotated Corpora (treebanks, etc.)	Parallel Multiling Corpora	Monoling. Lexicons	Multiling Lexicons	Grammars	Monoling. Thesauri, Ontologies Wordnets	Multiling. Thesauri, Ontologies Wordnets
Sentence Boundary Detection		+++	+				+		
Tokenizer	+++	+++	+						
Morphological Analyzer (deriv., stemm., diacritic, …)	+	+	+++		+++	+			
POS Tagger			+++		+++				
Chunker (Shallow Parser)			+++				+++		
Named Entity Recognizer			+++		+++	+++		+++	++
Word Sense Disambiguation					+++		+++		
Syntactic Analyzer			+++				+++		
Semantic Analyzer (incl. coreference resolution)			+++					+++	
Language Identifier				+++		+++	+++		
Translation (MT, query translation…)				+++		+++	++		+++

Table 2: Application vs. Modules

These matrices result from a consensus on which modules are needed to develop the applications and for each of these modules which language resources they require, as well as their relative importance.

The BLARK concept may be extended to specific languages or groups of languages in the specific field of MLIA, using the above mentioned matrices. In the next few years, it would be useful to conduct regular BLARK surveys for a selected set of less-resourced languages identified as key languages for the future development of MLIA applications (e.g. Eastern European languages or Indian languages).

BLARK surveys could also be conducted by focusing on a specific domain (e.g. resources and modules dedicated to multilingual information access in medical databases). This is an important first step to decide what specific resources should be developed and to prioritize such development in a specific context.

4.2.2 Recommendations

Along with the identification of language specific resource gaps; it is necessary to identify more precisely the resource needs in terms of domain specific data.

The underlying idea here is to get closer to the real needs of present and future potential CLIR system developers (who should consider the user needs), in order to help boost technology transfers from the labs to market. Research oriented to real user needs (such as the Intellectual Property track in CLEF 2009) seems like a good direction. A large amount of new resources will be needed to develop CLIR systems focusing on domain specific applications, in particular:

- New annotated multimedia and multilingual collections (in particular annotated image collections) to support the growing development of multimodal IR technologies in a multilingual context.
- New parallel web resources (for the development of Web IR applications) and multilingual web data collections (benchmarking). New language specific and domain specific web resources will be required within the next years (as was done with the EuroGOV corpus, which focussed on European institutional websites). Ideally, a cross-language web retrieval test collection should have documents in a wide variety of languages (and not be dominated by a particular language, i.e. English), focussing on a natural domain for multilingual web search (realistic scenario) and be of sufficient size (a few Terabytes of data at least).

Given the efforts required, before developing such resources, a first and necessary step is to identify clearly what sort of content is attractive for multilingual retrieval users. This has to be done by performing in-depth analysis of real user needs that can be implemented by regular surveys combined with other instruments.

Conducting such surveys is one major goal of the TrebleCLEF project. TrebleCLEF has organized workshops (System Developers Workshop 25, MLIA User Communities Workshop 26) that provided valuable information on present and future needs. Future reports on *System and User-oriented MLIA Best-practices* and *Best Practices for Test Collection Creation and Evaluation Methodologies* will provide information which may help to influence the production of resources in relevant directions.

Cross-disciplinary collaborations

Since the creation costs of complex multilingual and multimodal data collections are very high, building these resources is necessary a community effort, with each participant bringing in his/her particular language and other technical expertise, with a common identification of natural and realistic cross-language information needs.

As the complexity and variety of data is growing (more languages, more media, more particular domains), a successful future for the MLIA deployment rests more than ever upon collaborations across different research communities (private industry and academia), research fields (MLIA,

multimedia IR, web search, speech recognition, OCR, etc.) and user communities (medical, legal, humanities, publishing houses, etc.).

When addressing multimodal search engine challenges in particular, the development of MLIA should rely on closer collaborations with existing initiatives, such as Chorus[18] or Quaero[19] to name but a few. These collaborations are particularly critical with respect to the development of relevant benchmarking and evaluation resources. Due to the richness of the scientific objectives of multimedia search engines corresponding to growing and evolutionary use-cases and user needs, large collaborative benchmark initiatives are necessary to match the dynamicity of the field.

As far as multilinguality is concerned, large bilingual test corpora are urgently needed in order to evaluate and compare methods in a rational and scientific manner.

Distribution of resources

The valuable resources and experimental collections created by the major MLIA evaluation campaigns must be made available by promoting and coordinating the re-use, packaging and distribution of these data to the relevant communities. This is an important way to sustain an R&D community by providing high quality access to past evaluation results thus boosting the R&D activities through further dissemination of know-how, tools, resources and best practice guidelines. This is one of the objectives of the TrebleCLEF project, with respect to the CLEF evaluation resources.

The past activities emphasised the critical need for the evaluation packages (document collections, test queries, relevance judgments, and other commonly developed resources such as particular NLP tools etc.) to be made available after the evaluation campaigns.

It would also be interesting to search and identify relevant evaluation project in other scientific fields, which could provide interesting domain specific resources for the MLIA research.

Common pools of resources

Apart from the collaborative development for large data collections expressed by the respondents, there is a strong demand for a common development of other resources (lexicons, NLP tools etc.), especially among academic laboratories working with underrepresented languages. The MLIA community should improve its way of sharing resources using the services of well established and reliable data repositories. Very often, several groups work on the same target collections; these collections could be pre-processed in a joint community effort. In particular, it would be helpful to jointly develop common pools of resources and NLP tools for the underrepresented languages requiring a particular development effort which may be too high for a single centre.

A good example of this is what is being done with the Romanian language through ConsILR[20], the Consortium for the Romanian Language: Resources & Tools. ConsILR members can share resources and tools dedicated to Romanian language, as well as specialized publications and projects.

A more recent example is given by CLIA (Cross Lingual Information Access). The CLIA Project is a mission mode project funded by Government of India. It is being executed by a consortium of 11 academic and research institutions and industry partners. CLIA intends to develop in common a set of resources and document-processing modules to facilitate the access of documents written in English, Hindi and in many other Indian languages. Documents crawled from the web are pre-processed using

[18] Chorus: http://www.ist-chorus.org. Chorus is a FP6 Coordination Action on Advancing Search Technology for Audio-Visual Content. Chorus addresses this topic by putting together various initiatives on the benchmarking and evaluation of multimedia information retrieval to establish a clear understanding of the current situation and determine how best to move forward in a unified and cooperative way.

[19] Quaero : http://www.quaero.org. The main objective of Quaero is to develop applications corresponding to identified use cases in the domain of access and manipulation of multimedia and multilingual content.

[20] ConsILR, Consortium for the Romanian Language: Resources & Tools: http://consilr.info.uaic.ro/

language processing tools for extracting information and translating the extracted information into target languages. This module consists of many commonly developed NLP tools such as Document Converter, Language Pre-processors, POS taggers, Text Chunker, Named Entity Recognizer, Domain Dictionaries, Information Extraction Engine and Translation engine.

Other actions of this kind should be started for other less-represented languages.

5 Summing Up

Evaluation campaigns (e.g. TREC followed by NTCIR and CLEF) have proved crucial in stimulating research in IR and CLIR. In the beginning, these campaigns mostly focused on news corpora in textual form. Nowadays, there is a need for systems that are tuned to search efficiently within different types of data, such as web pages or scientific and technical corpora (i.e. patents, case law). Similarly, research used to be concentrated almost exclusively on text, with little attention to cross-media retrieval. There is now growing interest in functionality for search across language boundaries in multimedia collections

This report has carried out an inventory of existing Language Resources that can be exploited by MLIA developers and has identified a large number of basic tools and Language Resources that would be useful for the current and next generation of MLIA. The input for the report consisted of the results of surveys conducted on the groups that participate in the CLEF evaluation campaigns together with an extensive study of the literature. However, it is difficult (if not impossible) for a report of this type to be exhaustive. It should thus be considered as a useful beginning. It is our intention not only to publish the report on the TrebleCLEF website and to disseminate it as widely as possible among interested communities. We also intend to make an interactive version of the report available on the TrebleCLEF portal so that it can be dynamically added to and updated by the MLIA research community.

The findings reported here should be completed through consultations with language industry and communication players in order to draw a more precise picture of the technology developer's requirements with the corresponding assessed priorities. In this respect, the information and conclusions derived from the System Developers Workshop 25 and the MLIA User Communities Workshop 26 organized within the TrebleCLEF project constitute a good starting point.

We also report on a first attempt to design a BLARK for MLIA and provide an action plan and recommendations proposing the steps that need to be taken to develop the most critical set of language resources for all technologies related to MLIA and modules concerned by those technologies.

The action plan recommended here should lead to a roadmap that will outline areas and priorities for collaboration between technology developers, NLP developers and experts, language resources producers and in terms of collaboration between different disciplines to cross the barriers of languages and media/modalities.

The development of the roadmap will be based on a clear picture of the foreseeable technological trends, market potentials, and cooperation possibilities that can be drawn only from initiatives like CLEF (and TrebleCLEF) that federate a large number of community representatives. Such a roadmap will have to consider revising and updating the BLARK (the minimum set of resources and tools necessary for carrying out research and training on MLIA) on a regular basis. It is also of paramount importance that such BLARKs and roadmaps are endorsed by the community and obtain strong backing as the implementation of any recommendations made would be only achieved through a huge joint effort.

If research in the IR field is to be successful in the future, close collaboration across research communities (web search, speech recognition, OCR, text summarization, data mining, etc.) and user communities (medical, legal, humanities, etc.) will be crucial. In particular it seems necessary in the next years to pursue and further develop the CLEF activities towards more multimedia retrieval and application-oriented, domain specific tasks.

6 References

1. Deutsches Forschungszentrum für Künstliche Intelligenz (DFKI GmbH), *The ELSNET Roadmap for Human Language Technologies*, DFKI Report, November 2002.
2. Porter M. F. *An algorithm for suffix stripping*, Readings in Information Retrieval, Morgan Kaufmann Publishers Inc., pp. 313-316, 1997.
3. Eckhard B. *The Parsing System "Palavras": Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Framework.* Aarhus: Aarhus University Press, 2000.
4. Hartrumpf, S. *Coreference resolution with syntactico-semantic rules and corpus statistics*. In: Proceedings of the Fifth Computational Natural Language Learning Workshop (CoNLL-2001), pp. 137-144, Toulouse, France, 2001.
5. Hartrumpf S. *et al. The Semantically Based Computer Lexicon HaGenLex - Structure and Technological Environment***.** In: Traitement automatique des langues, 44(2), pp. 81-105, 2003.
6. Richter M. *et al*. *Exploiting the Leipzig Corpora Collection.* Proceedings of IS-LTC'06, Ljubljana, Slovenia, 2006.
7. Steinberger R. *et al. The JRC-Acquis: A multilingual aligned parallel corpus with 20+ languages*. Proceedings of LREC'2006, pp. 2142-2147. Genoa, Italy, 2006.
8. Koehn P., *Europarl: A Parallel Corpus for Statistical Machine Translation*, Proceedings of MT Summit 2005.
9. Grubinger M., Leung C., *A Benchmark for Performance Calibration in Visual Information Search*. In Proceedings of the 2003 International Conference on Visual Information Systems (VIS 2003), pp. 414–419, Miami, USA, 2003.
10. Cabrio, E. *et al.* The QALL-ME benchmark: a Multilingual Resource of Annotated Spoken Requests for Question Answering. *Proceedings of 6th Language Resources and Evaluation Conference (LREC2008)*, Marrakesh, Morocco, May 2008.
11. Buscaldi D. and Rosso P. Geo-wordnet: Automatic georeferencing of wordnet. *Proceedings of 6th Language Resources and Evaluation Conference (LREC2008)*, Marrakesh, Morocco, May 2008.
12. Briscoe E. and Carroll J. Robust accurate statistical annotation of general text. In *Proceedings of LREC'2002*, pp.1499-1504, Las Palmas, Gran Canaria, 2002.
13. Curran J.R., Clark S. and Bos J. Linguistically Motivated Large-Scale NLP with C&C and Boxer. *Proceedings of the ACL 2007 Demonstrations Session* (*ACL-07 demo*), pp.29-32, 2007.
14. Bos J., Clark S., Steedman M., Curran J.R. and Hockenmaier J. "*Wide-Coverage Semantic Representations from a CCG Parser*", Proceedings of the 20th International Conference on Computational Linguistics (COLING-04), pp.1240-1246, Geneva, Switzerland, 2004.
15. Och F. J., Ney H. "The alignment template approach to statistical machine translation." Computational Linguistics, volume 30, number 4, pages 417-449, 2004, MIT Press.
16. Cavnar W. B. and Trenkle J. M., ``N-Gram-Based Text Categorization" In Proceedings of Third Annual Symposium on Document Analysis and Information Retrieval, Las Vegas, NV, UNLV Publications/Reprographics, pp. 161-175, 11-13 April 1994.
17. Pradhan S., Ward W., Hacioglu K., Martin J., Jurafsky D. **"**Shallow Semantic Parsing using Support Vector Machines" in Proceedings of the Human Language Technology Conference (HLT/NAACL-2004), Boston, MA, May 2-7, 2004.
18. Erk K. and Pado S. "Shalmaneser - a flexible toolbox for semantic role assignment", in Proceedings of LREC 2006, Genoa, Italy, 2006.
19. Koehn P., Hoang H., Birch A., Callison-Burch C., Federico M., Bertoldi N., Cowan B., Shen W., Moran C., Zens R., Dyer C., Bojar O., Constantin A. and Herbstc E. "Moses: Open Source Toolkit for Statistical Machine Translation". Annual Meeting of the Association for Computational Linguistics (ACL), demonstration session, Prague, Czech Republic, June 2007.

20. Rogati M., Yang Y. “Resource selection for domain-specific cross-lingual IR Full text”, in Proceedings of the 27th ACM SIGIR conference, Sheffield, United Kingdom, pp. 154–161, 2004.
21. Krauwer S., *ELSNET and ELRA: A common past and a common future*, in ELRA Newsletter Vol. 3 N. 2. 1998.
22. Mapelli V., Choukri K. *Report on a (minimal) set of LRs to be made available for as many languages as possible, and map of the actual gaps*, ENABLER project internal report, Deliverable 5.1, 2003.
23. Binnenpoorte D, De Friend F., Sturm J., Daelemans W., Strik H., Cucchinari C. *A Field survey for Establishing Priorities in the Development of HLT Resources for Dutch*, in Proceedings LREC 2002, Las Palmas de Gran Canaria, Spain, 2002.
24. Krauwer S., Maegaard B, Choukri K., Damsgaard Jørgensen L. *Report on Basic Language Resource Kit (BLARK) for Arabic*, NEMLAR Report, 2004.
25. Braschler M., Clough P., *Bringing Multilingual Information Access to Operational Systems – TrebleCLEF System Developers Workshop*, Public report of the TrebleCLEF project (Deliverable 3.1), December 2008.
26. Gonzalo J., Peñas A., Verdejo F., Peters C., *Workshop on Best Practices for the Development of Multilingual Information Access Systems: the User Perspective (TrebleCLEF User Communities Workshop) - Report*, Public report of the TrebleCLEF project (Deliverable 3.2), December 2008.

Appendix A: Languages vs. Resources and Tools

A.1 Table Abbreviations

A.1.1 Language Code

Languages are represented by their 2 letter ISO language code (ISO-639-1[21]).

A.1.2 Availability

F	Free
OS	Open Source
OU	Online Use
C	Commercial licence
P	Available only to project partners or evaluation participants
ELDA	Distributed through the ELDA catalogue (http://catalog.elra.info/)
LDC	Distributed through the LDC catalogue (http://www.ldc.upenn.edu/Catalog/)

A.2 Resources by Language

A.2.1 Test Collections

Name		Lang	Description	Avail.
1	CLEF Collections	en, de, fr, it, es, eu, pt, nl, fi, sv, ru, cs, bg, ro, hu, fa	Monolingual, multilingual and multimodal corpora in most European Languages: Link: http://www.clef-campaign.org/	F, ELDA
2	TREC Collections	en, es, zh, ar, de, fr ,it	TREC mostly deals with information retrieval in English, but also had a few tracks involving other languages (TREC3-5: es / TREC5;6;9: zh / TREC6-8: de, fr, it / TREC10-11: ar). Link: http://trec.nist.gov/	F, ELDA, LDC
3	LT4eL Corpora	en, bu, nl, de, cs, pl, ro, pt, mt	Multilingual corpora in the domain of eLearning. POS and semantically annotated. The corpus varies among languages (only partially parallel). Link: http://www.lt4el.eue	P
4	CoNLL-2002-2003	es, nl, en, de	The CoNLL Conference created corpora tagged with named entity annotations. Link:http://www.cnts.ua.ac.be/conll2003/ner/ + http://www.cnts.ua.ac.be/conll2002/ner/ Languages: Spanish, Dutch, English, German	C, F
5	NTCIR Test Collections	en, zh, ja, ko	News corpora in English (Taiwan News, China Times English News, Hong Kong Standard, etc.) Other evaluation corpora: collections of patent application documents, web crawls… Link: http://research.nii.ac.jp/ntcir/data/data-en.html	F, P
6	FIRE	hi, bn, mr, ta, te, pa, ml, en	The FIRE website provides test data and other resources (stemmers, bilingual dictionaries, etc.) to the FIRE participants. The evaluations cover several Indian languages (Hindi, Bengali, Marathi, Tamil, Telugu, Punjabi, Malayalam) and English. Link: http://www.isical.ac.in/~fire/	P, F
7	MUC Test Collections	en	Availability: The MUC-6 and MUC-7 datasets are distributed through LDC. Datasets of the previous years are available completely free of charge. Link: http://www-nlpir.nist.gov/related_projects/muc/	LDC, F

[21] Official list of ISO-639: http://www.loc.gov/standards/iso639-2/php/code_list.php

8	TAC	en	The Text Analysis Conference (TAC) provides large test collections for: QA, Recognizing Textual Entailment (RTE Challenge), and Summarization… Link: http://www.nist.gov/tac/ Languages: English	F
9	Amaryllis	fr, en, it, es, de, pt	Text collections in French: news articles from "Le Monde"; plus titles and summaries of scientific articles. (+ multilingual text collections (6 languages: fr, en, it, es, de, pt) extracted from the MLCC parallel corpus). Link: http://www.inist.fr/accueil/profran.htm	ELDA
10	EQueR test collections	fr	News articles (Le Monde, Le Monde Diplomatique, and reports from the French Senate) and domain specific medical corpus (scientific articles and guidelines for good medical practice) Link: http://www.technolangue.net/article.php3?id_article=195	ELDA

A.2.2 Monolingual Corpora

Name		Lang	Description	Avail.
11	Leipzig Corpora	ca, da, nl, en, et, fi, fr, de, it, ja; ko, no, sr, sv, tr	A collection of large (non-parallel) corpora (data sources: either newspaper texts or texts randomly collected from the web.) Link: http://corpora.informatik.uni-leipzig.de	F
12	TIDES	en, zh, ar	TIDES developed resources for trans-lingual information processing (information detection, extraction, summarization and translation). Link: http://projects.ldc.upenn.edu/TIDES/	LDC
13	BNC	en	The British National Corpus (BNC) is a 100 million word collection of samples of written and spoken language from a wide range of sources (mostly British English). Link: http://www.natcorp.ox.ac.uk/	C
14	Web 1T 5-gram	en	Web 1T 5-gram, web data set, contributed by Google Inc., containing English word n-grams (from unigrams to 5-grams) and their observed frequency counts.	LDC
15	OTA	en	The Oxford Text Archive (OTA) offers many linguistic resources, mainly un-annotated text corpora (mostly British English). Link: http://ota.ahds.ac.uk/	F
16	NEGRA	de	NEGRA is a syntactically annotated corpus of German newspaper texts. Free for research: http://www.coli.uni-saarland.de/	F
17	TIGER Treebank	de	The TIGER Treebank consists of German newspaper text, semi-automatically POS-tagged and annotated with syntactic structure (+ morphological and lemma information for terminal nodes). Link: http://www.ims.uni-stuttgart.de/projekte/TIGER/TIGERCorpus/	F
18	CELEX	nl	CELEX Dutch lexical database	ELDA
19	TwNC	nl	The Twente News Corpus (TwNC) is a collection of text data for language model training (newspaper data, teletext subtitling of broadcast news shows, news data downloaded from the WWW). Contact : hltgroup@cs.utwente.nl	F
20	Alpino Dutch Treebank	nl	Dutch Treebank corpus Link: http://www.let.rug.nl/~vannoord/trees/	F
21	Italian Treebanks	it	Italian Treebanks: Italian Syntactic-Semantic Treebank (ISST) / Venice Italian Treebank (VIT).	ELDA
22	I-CAB	it	Italian Content Annotation Bank (I-CAB) is an annotated corpus consisting news stories taken from an Italian local newspaper. It is annotated with semantic information. Link: http://tcc.fbk.eu/projects/ontotext/icab.html	F

23	BulTreeBank	bg	HPSG-based Treebank of Bulgarian. Link: http://www.bultreebank.org/	F
24	BulTreeBank Frequency List	bg	BulTreeBank Frequency List: Link: http://www.bultreebank.org/resources/BTB-FreqList100000.zip	F
25	BulTreeBank Stopword List	bg	BulTreeBank Stopword List: Link: http://www.bultreebank.org/resources/BTB-StopWordList.zip	F
26	ROCO	ro	Romanian collection of news, with about 7 mil tokens, morpho-syntactically annotated and lemmatized. Available for research purposes upon request from RACAI Romania: http://www.racai.ro/	F
27	RO-corpus	ro	Romanian corpus of newspaper articles and two novels, 50 mil. words; no annotation. Available for research purposes upon request at: rada@cs.unt.edu	F
28	Bokr Russian Reference Corpus	ru	Bokr Russian Reference Corpus http://bokrcorpora.narod.ru/	under development
29	HANCO	ru	HANCO: The Helsinki annotated corpus of Russian texts http://www.slav.helsinki.fi/hanco/index_en.html	under development
30	Tübingen Russian Corpora	ru	Russian Corpora developed at the university of Tübingen (including the Uppsala Corpus and Corpus of Interviews). http://www.sfb441.uni-tuebingen.de/b1/korpora.html	F
31	Uppsala Russian Corpus	ru	The Uppsala Corpus consists of some 600 Russian texts with a total of one million words, equally divided between informative and literary prose. http://www.slaviska.uu.se/ryska/index.html	F
32	Leeds Russian Internet Corpus	ru	Russian Internet Corpus and Russian frequency lists. Links: http://corpus.leeds.ac.uk/ruscorpora.html and http://corpus.leeds.ac.uk/serge/frqlist/	F
33	RNC	ru	Russian National Corpus http://www.ruscorpora.ru/	OU
34	Russian Newspaper Corpus	ru	Russian Newspaper Corpus http://www.philol.msu.ru/~lex/corpus/	F
35	XX century's Basque Corpus	eu	Basque corpus XX century Link: http://www.euskaracorpusa.net/XXmendea/Konts_arrunta_fr.html	???
36	ZT Corpus	eu	Basque Corpus of Science and Technology Link: http://www.ztcorpusa.net/cgi-bin/kontsulta.py	???
37	Arabic Gigaword	ar	Arabic Gigaword	LDC
38	Arabic blogs dataset:	ar	12,000 Arabic blogs with over 120,300 posts, released by ILPS (University of Amsterdam). Link: http://ilps.science.uva.nl/resources/	F
39	NLEL Arabic Resources	ar	Resources from NLEL (Natural Language Engineering Lab) of the University of Valencia: • ANERCorp: A Corpus of more than 150,000 words annotated for Arabic NER. • ANERGazet: A collection of 3 Gazetteers in Arabic. • more than 11,000 Arabic Wikipedia Articles in SGML format (the format adopted in the CLEF and also the one accepted by the JIRS system). • Arabic Wikipedia XML corpus Link: http://www.dsic.upv.es/grupos/nle/downloads.html	F
40	Hamshahri Corpus	fa	Corpus of text from the *Hamshahri* newspaper. Link: http://ece.ut.ac.ir/dbrg/hamshahri	F

41	Bijankhan Corpus	fa	The Bijankhan Corpus is a tagged corpus (daily news and common texts). Links: http://ece.ut.ac.ir/DBRG/Bijankhan/	F
42	LCMC	zh	The Lancaster Corpus of Mandarin Chinese (LCMC)	ELDA
43	Tagged Chinese Gigaword	zh	Tagged Chinese Gigaword, annotated with full part of speech tags.	LDC
44	ACLCLP Corpora	zh	Corpus Program (News Corpus), Sinica Balanced Corpus, and several speech corpora distributed by the Association for Computational Linguistics and Chinese Language Processing (ACLCLP). http://www.aclclp.org.tw/corp.php	C
45	Sinica Treebank	zh	Chinese Electronic Dictionary distributed by the ACLCLP. http://www.aclclp.org.tw/corp.php	C
46	CIRB030	zh	Chinese Information Retrieval Benchmark: Test collection to evaluate the performance of Chinese IR systems distributed by the ACLCLP. http://www.aclclp.org.tw/corp.php	C
47	Talkbank Korean Corpus	ko	Morphologically Annotated Korean Text	LDC
48	Korean Treebank and PropBank	ko	Korean Treebank + Korean Propbank (semantic annotation of the Korean English Treebank)	LDC
49	EMILLE Lancaster Corpus	hi, bn, gu, pa, si, ta, ur	EMILLE Lancaster Corpus: monolingual corpora for 7 South Asian languages (Bengali, Gujarati, Hindi, Punjabi, Sinhala, Tamil, Urdu.)	ELDA
50	EMILLE CIIL Corpus	hi, bn, gu, pa, si, ta, ur, as, kn, ks, ml, mr, or, te	EMILLE/CIIL Corpus: monolingual corpora for 14 South Asian languages (Assamese, Bengali, Gujarati, Hindi, Kannada, Kashmiri, Malayalam, Marathi, Oriya, Punjabi, Sinhala, Tamil, Telegu and Urdu.)	ELDA
51	ILPS Indonesian Resources	id	Corpora and other IR resources for Bahasa Indonesia, released by ILPS (University of Amsterdam). Link: http://ilps.science.uva.nl/resources/	F
52	Kurdish Resources	ku	Corpora and other r resources for Kurdish. Link: http://www.cogsci.ed.ac.uk/~siamakr/kurd_lal.html	F

A.2.3 Multilingual, Parallel Corpora

Name		Lang	Description	Avail.
53	Treebanks	en, ar, zh, ko	Monolingual English Treebank + bilingual treebanks (English-Arabic, -Chinese, -Korean…)	LDC
54	JRC-Acquis	bg, cs, da, de, el, en, es, et, fi, fr, hu, it, lt, lv, mt, nl, pl, pt, ro, sk, sl, sv	Aligned parallel corpora covering 22 official EU languages (EU legislative texts). Link: http://langtech.jrc.it/JRC-Acquis.html	F
55	Europarl	en *vs.* el, es, de, da, fi, fr, it, nl, pt, sv	European Parliament Proceedings Parallel aligned corpora covering 11 official EU languages (10 languages aligned to English) extracted from the proceedings of the European Parliament. Link: http://www.statmt.org/europarl/.	F
56	QALL-ME	en *vs.* it, es, de	The QALL-ME corpus is a multilingual resource of annotated spoken requests in the tourism domain (Annotations: pragmatic, semantic and QA-based annotations). Link: http://qallme.fbk.eu/	F

57	MULTEXT-JOC	en, fr, de, it, es	JOC MULTEXT: raw, grammatically tagged and aligned data from the *Written Questions and Answers* of the Official Journal of the European Community. Language: English, French, German, Italian and Spanish.	ELDA
58	ARCADE / ROMANSEVAL	en, fr, it, es, pt	ARCADE / ROMANSEVAL: raw and annotated data from the JOC corpus.	ELDA
59	MULTEXT-East	en, bg, hr, cs, et, hu, lt, ro, ru, sl, sr	Parallel and comparable corpora in English and Eastern European languages: Bulgarian, Croatian, Czech, Estonian, Hungarian, Lithuanian, Romanian, Russian, Slovene, and Serbian (text corpora under license, speech corpora are free). Link: http://nl.ijs.si/ME/	F, C
60	MuchMore Springer Corpus	en, de	English-German scientific medical abstracts aligned on the sentence level (annotation: POS tags; morphology semantic classes, semantic relations…) Link: http://muchmore.dfki.de/resources_index.htm	F
61	GALE Corpora	en, ar, zh	GALE Arabic(-English) Broadcast News Parallel Text GALE Chinese(-English) Broadcast News Parallel Text	LDC
62	ISI corpora	en, zh	ISI Chinese-English Automatically Extracted Parallel Text.	LDC
63	TC-STAR	en, es, zh	TCSTAR corpora (English, Spanish, Chinese)	ELDA
64	CESTA Corpora	fr, en, ar	CESTA corpora consist of several English-French and Arabic-French parallel corpora.	ELDA
65	EMILLE CIIL & Lancaster Corpus	en *vs.* hi, bn, pa, gu, ur	Contains parallel corpora of text in English with translations in 5 Indian languages (Hindi, Bengali, Punjabi, Gujarati and Urdu.)	ELDA
66	MLCC corpora	da, nl, en, fr, de, el, it, pt, es	MLCC corpora include monolingual corpora of newspaper articles in 6 European languages (Dutch, English, French, German, Italian and Spanish.) and a multilingual parallel corpus consisting of translated data in 9 European languages.	ELDA
67	UN Parallel Corpora	en, fr, es	UN parallel text corpora.	LDC
68	Canadian Hansard	en, fr	Canadian Hansard parallel corpora (English vs. French).	LDC
69	Hong Kong Parallel Text	en, zh	Hong Kong Parallel Text: Hong Kong Laws, Hong Kong News, and Hong Kong Hansards parallel text corpora (English vs. Chinese.)	LDC
70	Kaist Korterm Corpora	en, zh, ko	Kaist Korterm multilingual parallel corpus: 60 000 expressions in Korean, Chinese and English.	ELDA
71	MultiSemCor	en, it	MultiSemCor is an English/Italian parallel corpus, aligned at the word level and annotated with PoS, lemma and word sense. Link http://multisemcor.itc.it/	F
72	MD Corpora	fr *vs.* ar, zh, el, ja, fa, ru	The MD corpora (Le Monde Diplomatique) contain texts in Arabic, Chinese, Greek, Japanese, Persian and Russian manually aligned with French. A subset for the Arabic French part was manually annotated with named entities.	ELDA
73	Prague Dependency Treebank	en, cs	Czech-English Treebank	LDC
74	1984 English-Romanian Corpus	en, ro	Orwell's novel as a parallel English-Romanian corpus, annotated for POS, lemma and chunks, with word alignments. For research purposes – available upon request from RACAI Romania: http://www.racai.ro/	F
75	Republica	fr, ro	Plato's *Republic*: French-Romanian parallel corpus with about 250 thousand tokens, morpho-syntactically annotated. For research purposes – available upon request from RACAI Romania: http://www.racai.ro/	F

76	NAACL English-Romanian Corpora	en, ro	NAACL'03 Training Data: English-Romanian parallel text of *1984*, Romanian Constitution, and a large (about 900,000 tokens) collection of texts collected from the Web; sentence-aligned. For research purposes – available upon request from rada@cs.unt.edu NAACL-News '03: Romanian-English word aligned data, morpho-syntactically annotated and lemmatized. http://www.cs.unt.edu/~rada/wpt/data/Romanian-English.test.tar.gz NAACL-News '05: Collection of news, with about 850.000 tokens, with XML annotations for POS, lemma, chunks, word alignments. http://www.cs.unt.edu/~rada/wpt05/data/Romanian-English.test.tar.gz Link: http://www.cs.unt.edu/~rada/wpt	F
77	RoTimeBank	en, ro	Parallel, word-aligned English-Romanian corpus (about 63.000 tokens each side) with the translations of the English original corpus; annotations for POS, lemma, chunks, TimeML entities, NE (MUC style) Romanian part available for research purposes upon request from corinfor@info.uaic.ro (as of Feb 2009; soon available on web)	F
78	RoFrameNet Corpus	en, ro	1094 sentences (25.000 tokens) translated from the English FrameNet, with annotations for POS, lemma, semantic roles. For research purposes – available upon request from corinfor@info.uaic.ro	F
79	EMILLE CIIL & Lancaster Corpus	en *vs.* hi, bn, pa, gu, ur	Parallel corpora of text in English with translations in 5 Indian languages (Hindi, Bengali, Punjabi, Gujarati and Urdu.)	ELDA

A.2.4 Multimodal Corpora

Name		Lang	Description	Avail.
80	IAPR TC-12	en, de, es	Collection of still natural images. Each image is associated with a text caption in up to three different languages (English, German and Spanish). Link: http://eureka.vu.edu.au/~grubinger/IAPR/TC12_Benchmark.html	F
81	INEX Corpora	en, de, fr, nl, es, zh, ar, ja	Initiative for the Evaluation of XML Retrieval (INEX): Wikipedia XML collection, Wikipedia image collection Link: http://inex.is.informatik.uni-duisburg.de	P
82	TRECVID Corpora	en, zh, ar, nl	Corpora (video, key frames, transcripts) for content-based video retrieval. Link: http://www-nlpir.nist.gov/projects/trecvid/	LDC
83	CHIL Corpora	en	Corpora (video, transcripts) of audio-visual recordings of meetings.	ELDA
84	FarsDat	fa	This Persian speech database includes the recordings of 300 Persian speakers spanning several dialects. The data has been segmented and phonemically labelled. Link: http://www.elda.org/catalogue/en/speech/S0112.html	ELDA
85	CallFriend	fa	60 unscripted telephone conversations in Farsi.	LDC

A.2.5 Monolingual Lexica and Dictionaries

Name		Lang	Description	Avail.
86	Leipzig Dictionaries	48 languages	Online search in 48 Corpus-Based Monolingual Dictionaries (48 languages). Link: http://corpora.informatik.uni-leipzig.de	F
87	MULTEXT Lexicons	en, fr, de, it, es	Lexicons developed in the MULTEXT project.	ELDA
88	Monolingual SCIPER dictionaries	en, de, fr, es, it	Monolingual dictionaries produced in the frame of the EURADIC project.	ELDA

89	LDOCE Dictionary	en	Longman Dictionary of Contemporary English Link: http://www.ldoceonline.com/	C
90	BESL	en	British English Source Lexicon (BESL).	ELDA
91	WEB-DEX	ro	Romanian explanatory dictionary, XML-encoded. For research purposes – available upon request from RACAI Romania: : http://www.racai.ro/	F
92	RO-lex	ro	Romanian lexicon with about 580000 entries, including: wordform, lemma, POS, etc. For research purposes – available upon request from RACAI Romania: : http://www.racai.ro/	F
93	Russicon Resources	ru	Dictionaries and thesauri developed by the Russicon company. http://www.russicon.ru/all.htm http://schools.keldysh.ru/uvk1838/Sciper/volume2/langres/russiclr.htm	C
94	EDBL	eu	EDBL (Euskararen Datu-Base Lexikala) is a lexical database for Basque. http://ixa2.si.ehu.es/cgi-bin/mcr/public/wei.consult.perl	ELDA
95	Chinese Electronic Dictionary	zh	Chinese Electronic Dictionary distributed by the ACLCLP. http://www.aclclp.org.tw/corp.php	C

A.2.6 Multilingual Lexica and Dictionaries

Name		Lang	Description	Avail.
96	Bilingual EURADIC dictionaries	fr, en, de, es, it	Bilingual dictionaries: French-English, French-German, French-Spanish, French-Italian, English-German, English-Spanish.	ELDA
97	CJK Dictionary Institute	en, zh, ja, ko	Multilingual dictionaries: English, Chinese, Japanese, Korean. Link: http://www.cjk.org/	C
98	Chinese-English Translation Lexicon	zh, en	The GALE project developed several statistical translation lexicons, including the Chinese-English lexicon.	LDC
99	LT4eL Lexicons	bg, en, nl, de, cs, pl, ro, pt	Multilingual lexicons in the domain of eLearning, covering the concepts in the LT4eL ontology (Bulgarian, English, Dutch, German, Czech, Polish, Romanian, Portuguese). Link: http://www.lt4el.eu	F
100	DixAF	fr, ar	DixAF is a bilingual French-Arabic dictionary developed at the CNRS.	ELDA
101	FR-RO dictionary	fr, ro	French Romanian Dictionary: 16,710 entries, XML – implemented to be compatible with the TEI-light specifications for bilingual dictionaries. For research purposes – available upon request from RACAI Romania: http://www.racai.ro/	F
102	EN-RO dictionary	en, ro	English-Romanian Dictionary: 38,000 entries, XML. Link: http://lit.csci.unt.edu/~rada/downloads/RoNLP/R.E.tralex	F
103	GerLexicon	en, de, ro	German, English, and Romanian Lexicons, with bilingual connections between them. Link: http://nats-www.informatik.uni-hamburg.de/view/Main/GerLexicon	F
104	En-Ru SCIPER dictionary	en, ru	Bilingual Russian-English SCIPER dictionary	ELDA
105	Shiraz Persian to English Dictionary	en, fa	The Shiraz Persian-to-English dictionary consists of approximately 50,000 entries including single words, phrases and proper names. Link: http://crl.nmsu.edu/Resources/lang_res/persian.html	F
106	Persian Lexicon Project	en, fa	Persian-English lexicon http://crl.nmsu.edu/Resources/lang_res/persian.html	F
107	CDICT	en, zh, ja	Chinese-English-Japanese Dictionary. Link: http://cdict.freetcp.com/	OU
108	CC-CEDICT	zh, en	Freely available Chinese to English dictionary. Link: http://us.mdbg.net/chindict/chindict.php	F

A.2.7 Monolingual Ontologies, Thesauri and WordNets

Name	Lang	Description	Avail.
109 Getty TGN	en	Getty Thesaurus of Geographic Names (TGN). Online use: http://www.getty.edu/research/conducting_research/vocabularies/tgn/	C
110 Getty AAT	en	Getty Art & Architecture Thesaurus (AAT). Online use: http://www.getty.edu/research/conducting_research/vocabularies/aat/	C
111 Getty ULAN	en	Getty Union List of Artists Names (ULAN). Online use: http://www.getty.edu/research/conducting_research/vocabularies/ulan/	C
112 GeoWordNet	en	GeoWordNet is a geo-referenced version of WordNet. http://www.dsic.upv.es/grupos/nle/resources/geo-wn/download.html	F
113 English FrameNet	en	FrameNet lexical database of lexical units (word + meaning) and semantic frames. Link: http://framenet.icsi.berkeley.edu/	C, F
114 Roget's Thesaurus	en	Roget's Thesaurus for English, version 1911. Link: http://machaut.uchicago.edu/rogets	F
115 WordNet	en	Large lexical database of English + semantic annotations. Link: http://wordnet.princeton.edu/	F
116 German FrameNet	de	FrameNet lexical database of lexical units (word + meaning) and semantic frames. Link: http://gframenet.gmc.utexas.edu	C, F
117 GermaNet	de	German version of WordNet Link: http://www.sfs.uni-tuebingen.de/GermaNet/	C, F
118 IMSLex	de	IMSLex is a lexicon for German with morphosyntactic and subcategorization information. Released by IMS (Univ of Stuttgart): http://www.ims.uni-stuttgart.de/projekte/corplex/	F
119 OpenThesaurus	de	OpenThesaurus is a free German thesaurus and WordNet. Link: http://www.openthesaurus.de	F
120 WOLF	fr	WordNet Libre du Français (Free French WordNet) Link: http://alpage.inria.fr/~sagot/wolf-en.html	F
121 Spanish FrameNet	es	FrameNet lexical database of lexical units (word + meaning) and semantic frames. Link: http://gemini.uab.es/SFN/	C, F
122 WordNet.PT	pt	Portuguese WordNet. Link: http://www.clul.ul.pt/wn2/	OU
123 GeoNET-PT	pt	GeoNET-PT is a geographical ontology in Portuguese. Link: http://poloxldb.linguateca.pt/index.php?l=geonetpt	F
124 Lingateca Resources	pt	Linguateca (distributed language resource centre for Portuguese) is the reference resource repository for Portuguese resources. It includes ontologies, thesauri, lexica, etc. Link: http://www.linguateca.pt/	F
125 Greek Wordnet (BalkaNet)	el	Seen in the list of available WordNets in the world provided by the Global WordNet Association: http://www.globalwordnet.org/	???
126 DanNet (Danish WordNet)	da	Seen in the list of available WordNets in the world provided by the Global WordNet Association: http://www.globalwordnet.org/	???
127 Swedish WordNets	sv	2 Swedish WordNets seen in the list provided by the Global WordNet Association: http://www.globalwordnet.org/	???
128 Maltese WordNet	mt	Seen in the list of available WordNets in the world provided by the Global WordNet Association: http://www.globalwordnet.org/	???
129 Polish WordNet	pl	Polish WordNet seen in the list of available WordNets in the world provided by the Global WordNet Association: http://www.globalwordnet.org/	???
130 plWordNet	pl	Polish WordNet seen in the list of available WordNets in the world provided by the Global WordNet Association: http://www.globalwordnet.org/	???

131	sloWNet	sl	Slovenian WordNet Link: http://lojze.lugos.si/~darja/slownet.html	F
132	BulNet	bg	Bulgarian WordNet	ELDA
133	BalkaNet	bg	See the list of available WordNets in the world provided by the Global WordNet Association: http://www.globalwordnet.org/	???
134	RoWordNet	ro	Romanian WordNet. Link: http://nlp.racai.ro/wnbrowser/	OU
135	Romanian WordNet (BalkaNet)	ro	Seen in the list of available WordNets in the world provided by the Global WordNet Association: http://www.globalwordnet.org/ Web service: http://nlp.racai.ro/wnbrowser/, and available for research purposes upon request from RACAI Romania: http://www.racai.ro/	OU, F
136	ConsILR Pool of Resources	ro	The ConsILR (Consortium for the Romanian Language: Resources & Tools) offers resources for Romanian: annotated corpora, thesauri, dictionaries Link: http://consilr.info.uaic.ro/ Mostly restricted to ConsILR members.	P
137	Hungarian Wordnet	hu	Seen in the list of available WordNets in the world provided by the Global WordNet Association: http://www.globalwordnet.org/	???
138	RussNet	ru	Russian WordNet Link: http://www.philarts.spbu.ru/depts/12/RN/	C
139	Russian WordNets	ru	2 other Russian WordNets seen in the list provided by the Global WordNet Association: http://www.globalwordnet.org/	F
140	AlbaNet	sq	See the list of available WordNets in the world provided by the Global WordNet Association: http://www.globalwordnet.org/	???
141	Norwegian WordNet	no	Seen in the list of available WordNets in the world provided by the Global WordNet Association: http://www.globalwordnet.org/	???
142	EuroWordNet for Basque	eu	EuroWordNet for Basque Link: http://ixa2.si.ehu.es/cgi-bin/mcr/public/wei.consult.perl	???
143	Serbian Wordnet (BalkaNet)	sr	Seen in the list of available WordNets in the world provided by the Global WordNet Association: http://www.globalwordnet.org/	???
144	Croatian WordNet	hr	Seen in the list of available WordNets in the world provided by the Global WordNet Association: http://www.globalwordnet.org/ Link: http://rmjt.ffzg.hr/p3_en.html	???
145	Turkish Wordnet (BalkaNet)	tr	Seen in the list of available WordNets in the world provided by the Global WordNet Association: http://www.globalwordnet.org/	???
146	Arabic WordNet	ar	See the list of available WordNets in the world provided by the Global WordNet Association: http://www.globalwordnet.org/	F
147	Hebrew WordNet	he	Hebrew version of WordNet. Link: http://cl.haifa.ac.il/projects/mwn/	F
148	PersiaNet	fa	Persian WordNet seen in the list of available WordNets in the world provided by the Global WordNet Association: http://www.globalwordnet.org/	???
149	FarsNet	fa	Persian WordNet seen in the list of available WordNets in the world provided by the Global WordNet Association: http://www.globalwordnet.org/	???
150	Chinese WordNet	zh	Seen in list of available WordNets provided by the Global WordNet Association: http://www.globalwordnet.org/ & http://bow.sinica.edu.tw/	???
151	Japanese FrameNet	ja	FrameNet lexical database of lexical units (word + meaning) and semantic frames. Link: http://jfn.st.hc.keio.ac.jp/	C, F
152	GoiTaikei	ja	Japanese dictionary marked using a semantic ontology. Link: http://www.kecl.ntt.co.jp/mtg/resources/GoiTaikei	C

153	Japanese Wordnet	ja	See the list of available WordNets in the world provided by the Global WordNet Association: http://www.globalwordnet.org/	???
154	Korean WordNets	ko	KorLex, AlexKor, …(seen in the list of available WordNets in the world provided by the Global WordNet Association: http://www.globalwordnet.org/)	???
155	Hindi WordNet	hi	Hindi version of WordNet.	LDC
156	Indian language WordNets	ta, sa, *etc.*	WordNets of several Indian languages (Tamil, Sanskrit, etc.) seen in the list provided by the Global WordNet Association: http://www.globalwordnet.org/	???

A.2.8 Multilingual Ontologies, Thesauri and WordNets

Name		Lang	Description	Avail.
157	UMLS Metathesaurus	cs, da, de, es, en, eu, fi, fr, he, hu, it, nl, no, pt, ru, sv	The UMLS thesaurus is developed by the National Library of Medicine (NLM). Link: http://www.nlm.nih.gov/pubs/factsheets/umlsmeta.html	F
158	MeSH	en *vs.* de, fr, es, pt, it, fi, ru	National Library of Medicine's thesaurus of Medical Subject Headings (English and translations in the 7 other languages). Link: http://www.nlm.nih.gov/mesh/	F
159	EuroWordNet	en, nl, it, es, de, fr, cs, et	EuroWordNet is a multilingual database with wordnets for several European languages. Link: http://www.illc.uva.nl/EuroWordNet/	ELDA
160	MLSN	en, de, ja, zh	Multi-Lingual Semantic Network Link: http://dcook.org/mlsn/	F
161	LT4eL Ontology	bg, en, nl, de, cs, pl, ro, pt	Ontology in the domain of information technology for end users. Link: http://www.lt4el.eu	F
162	HaGenLex / HaEnLex	de, en	HaGenLex (Hagen German Lexicon): domain independent computational lexicon. (morphosyntactic and semantic annotations). HaEnLex is the English version. Link: http://pi7.fernuni-hagen.de/forschung/hagenlex/hagenlex-en.html	C
163	MultiWordNet	en, it	Multilingual lexical database: Italian WordNet aligned on Princeton's WordNet. Link: http://multiwordnet.fbk.eu	C, F
164	Sinica Bilingual WordNet	zh, en	Mandarin-English bilingual database distributed by the ACLCLP. http://www.aclclp.org.tw/corp.php	C
165	Sinica Bilingual Ontological Database	zh, en	Chinese-English bilingual ontological database distributed by the ACLCLP. http://www.aclclp.org.tw/corp.php	C
166	AWN	ja, ko, th, vi, bn, mn	Asian WordNet is an interconnected WordNet for Asian languages: Thai, Korean, Japanese, Indonesian, Myanmar, Vietnamese, Mongolian, Bengali. Link: http://asianwordnet.org/	F
167	CLIA Resources	hi, *etc.*	The Cross Language Information Access (CLIA) consortium for Indian languages produces resources: Wordnet for Hindi (IIIT Bombay), bilingual dictionaries for various pairs of Indian languages, Language Models created from monolingual corpora, parallel corpus of named entities among various Indian languages…	P

A.2.9 Languages vs. Resources

The numbers in the table below refer to the identifying numbers of the above mentioned resources.
Abbreviations:

Test-Coll	Test Collections
Mono-C	Monolingual Corpora
ML-C	Multilingual Parallel Corpora
MM-C	Multimodal Corpora
Mono-L	Monolingual Lexica, Dictionaries
ML-L	Multilingual Lexica, Dictionaries
Mono-O	Monolingual Ontologies, Thesauri, WordNets etc.
ML-O	Multilingual Ontologies, Thesauri, WordNets etc.

Lang.	Test-Coll	Mono-C	ML-C	MM-C	Mono-L	ML-L	Mono-O	ML-O
en	1-8	11, 12, 13, 14, 15	53-71, 73, 74, 76, 77, 78	80-83	86-90	96-99, 102, 103, 104	109-115	157-163
de	1, 2, 3, 4	11, 16, 17	54, 55, 56, 57, 60, 66	80, 81	86, 88	96, 99, 103	117, 118, 119	157, 159, 160, 158, 161, 162
nl	1, 3, 4	11, 18, 19, 20	54, 55, 66	81, 82	86	99		157, 159, 161
fr	1, 2, 9, 10	11	54, 55, 57, 58, 64, 66, 67, 68, 72,	81	86, 88	96, 100, 101,	120	157, 158, 159
it	1, 2	11, 21, 22	54, 55, 56, 57, 58, 66, 71		86, 88	96		157, 158, 159, 163
es	1, 2, 4		54, 55, 56, 57, 58, 63, 66, 67	80, 81	86, 88	96	121	157, 158, 159
pt	1, 3		54, 55, 58, 66		86	99	122, 123, 124	157, 158, 161
el			54, 66, 72		86		125	
da		11	54, 55, 66		86		126	157
sv	1	11	54, 55		86		127	157
fi	1	11	54, 55		86			157, 158s
mt	3		54				128	
pl	3		54			99	129, 130	161
cs	1, 3		54, 59, 73		86	99		157, 159, 161
sk			54		86			
sl			54, 59		86		131	
bg	1, 3	23, 24, 25	54, 59		86	99	132, 133	161
ro	1, 3	26, 27	54, 59, 74, 76, 77, 78		86, 91, 92	99, 101, 102, 103	134, 135, 136	161
hu	1		54		86		137	157
et		11	54, 59		86			159
lt			54, 59		86			
lv			54		86			

Lang.	Test-Coll	Mono-C	ML-C	MM-C	Mono-L	ML-L	Mono-O	ML-O
ru	1	28, 29, 30, 31, 32, 33, 34	59, 72		86, 93	104	138, 139	157, 158
sq							140	
no		11			86		141	157
ca		11			86			
eu	1	35, 36			94		142	157
sr		11	59		86		143	
hr			59		86		144	
tr		11			86		145	
ar	2	12, 37, 38, 39	53, 61, 61, 64	81, 82			146	
he							147	157
fa	1	40, 41	72	84, 85		105, 106	148, 149	
zh	2, 5	12, 42, 43, 44, 45, 46	53, 61, 62, 63, 69, 70, 72	81, 82	86, 95	97, 98, 107, 108	150	160, 164, 165
ja	5	11	72	81	86	97, 107	151, 152, 153	160, 166
ko	5	11, 47, 48	70		86	97	154	166
hi[22]	6	49, 50	79		86		155, 156	166, 167
A+[23]		51, 52						166

[22] Hindi and other Indian languages.

[23] Other Asian languages: Indonesian, Vietnamese, Mongolian, Kurdish, …

A.3 Tools by Language

A.3.1 Stemmers and Morphologic Analysis

Name	Lang	Description	Avail.
1 UNINE Stemmers	en, de, fr, it, es, pt, sv, fi, pl, cs, bg, ro, hu, ru, ar, fa	Open source stemming algorithms (and stop-word lists). Link: http://members.unine.ch/jacques.savoy/clef/index.html	OS
2 Snowball stemmers	en, fr, es, pt, it, ro, de, nl, sv, no, da, ru, fi, hu, tr	Several stemmers including the Porter's stemmer, for English. Link: http://snowball.tartarus.org	F
3 RSLP Stemmer	pt	Portuguese stemmer developed by UFRGS (Universidade Federal do Rio Grande do Sul). Link: http://www.inf.ufrgs.br/%7Earcoelho/rslp/integrando_rslp.html	F
4 JSPELL	pt	Open source morphological analyzer "jspell". Link: http://natura.di.uminho.pt/wiki/doku.php?id=ferramentas:jspell	F
5 PALAVRAS	pt	Syntactic analyzer PALAVRAS in Portuguese. Link: http://beta.visl.sdu.dk/constraint_grammar.html	C
6 Czech parser and tagger	cs	A series software tools (Czech morphological analyzer and tagger).	LDC
7 ILPS Resources	hu	Resources for stemming and stopping in Hungarian released by ILPS (University of Amsterdam) Link: http://ilps.science.uva.nl/resources/	F
8 TTL (Tokenizer, Tagger and Lemmatizer)	en, ro	PERL module for text segmentation at sentence/word level, morpho-syntactic annotation and lemmatization. It is language independent and it was trained and used mainly for English and Romanian. Link: http://nlp.racai.ro/RacaiXmlWebServices.aspx	OU
9 DIAC+	ro	DIAC+ is an integrated system for automatically recovering of missing diacritics in Romanian texts. With adequate resources (a morpho-lexical disambiguated text, containing diacritics), DIAC+ can be used for other languages as well. Link: http://nlp.racai.ro/webservices/TextProcessing.aspx	OU
10 Ro-Hyphenator	ro	PERL application that performs automatic hyphenation for Romanian. Upon request (for research purposes) from RACAI Romania: http://nlp.racai.ro/	F
11 Morphological Analyser for Basque	eu	Morphological analyzer/generator for Basque Link: http://ixa2.si.ehu.es/demo/analisianali.jsp	OU
12 EusTagger	eu	Lemmatizer/Tagger for Basque Link: http://ixa2.si.ehu.es/demo/analisimorf.jsp	OU
13 Zatiak	eu	Chunker for Basque Link: http://ixa2.si.ehu.es/demo/zatiak.jsp	OU
14 Buckwalter Arabic Morphological Analyzer	ar	Buckwalter Arabic Morphological Analyzer.	LDC
15 Persian Morphological Analyzer	fa, ar, ur	CRL's Morphological Analyzer generates analyses for texts in Arabic, Persian and Urdu. Link: http://crl.nmsu.edu/Resources/lang_res/persian.html	F
16 Perstem	fa	Persian Stemmer and Morphological Analyzer http://students.cs.byu.edu/~jonsafar/	OS

Name	Lang	Description	Avail.
17 Stanford Chinese Word Segmenter	zh	Stanford Chinese Word Segmenter. Link: http://nlp.stanford.edu/software/segmenter.shtml	F
18 Chinese Segmenter	zh	This tool breaks a Chinese text file into words. Link: http://www.mandarintools.com/segmenter.html	F
19 ChaSen	ja	ChaSen is a morphological parser for the Japanese language. Link: http://chasen.naist.jp/hiki/ChaSen/	OS

A.3.2 POS Taggers

Name	Lang	Description	Avail.
20 Stanford POS Tagger	en, de, ar, zh	POS taggers from the Stanford NLP toolkits Link: http://nlp.stanford.edu/software/	F
21 TreeTagger	en, de, fr, it, nl, es, bg, ru, el, pt, zh	The TreeTagger is a POS-tagger developed within the TC project. Link: http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger	F
22 SVMTool	en, es, ca	SVMTool provides an open source POS tagger. Link: http://www.lsi.upc.edu/~nlp/SVMTool/	F
23 TnT POS Taggers	en, de	The TnT tagger is a statistical part-of-speech tagger. Link: http://www.coli.uni-saarland.de/~thorsten/tnt/	F
24 QTag	en, de	QTAG is a probabilistic POS tagger. In principle language independent, it is released only with resource files for English and German (software to create resources for other languages is included in the distribution.) Link: http://morphix-nlp.berlios.de/manual/node17.html	F
25 CLAWS	en	CLAWS (Constituent Likelihood Automatic Word-tagging System) is a POS tagging software for English text. Link: http://ucrel.lancs.ac.uk/claws/	C
26 RFTagger	de, cs, hu	The RFTagger is a tool for part-of-speech tagging. Link: http://www.ims.uni-stuttgart.de/projekte/corplex/RFTagger/	OS
27 Bulgarian Taggers	bg	Bulgarian Taggers: http://www.bultreebank.org/taggers/taggers.html	F

A.3.3 Syntactic Parsers

Name	Lang	Description	Avail.
28 Stanford Parsers	en, de, zh, ar	Syntactic parsers from the Stanford NLP toolkits Link: http://nlp.stanford.edu/software/	F
29 Machinese Syntax	en, fr, de, es, it, nl, sv, da, no, fi	Machinese Syntax is the syntactic parser of the Machinese tool suite commercialized by the Connexor Company. Link: http://www.connexor.eu/	C
30 Charniak Parsers	en	Statistical parsers developed by Eugene Charniak at the Brown University. Link: ftp://ftp.cs.brown.edu/pub/nlparser/	F
31 MINIPAR	en	MINIPAR is a broad-coverage parser for the English language. Link: http://www.cs.ualberta.ca/~lindek/minipar.htm	F
32 C&C Parser	en	The C&C Tools suite include the C&C CCG parser. Link: http://svn.ask.it.usyd.edu.au/trac/candc/wiki	F
33 RADISP	en	The RADISP is a domain-independent syntactic parser for English. Link: http://www.informatics.susx.ac.uk/research/groups/nlp/rasp/	F

Name	Lang	Description	Avail.
34 Link Grammar Parser	en	The Link Grammar Parser is a syntactic parser of English, based on link grammar, an original theory of English syntax. http://www.link.cs.cmu.edu/link	OS
35 Alpino Parser	nl	The Alpino parser is an open-source syntactic dependency parser for Dutch. Link: http://www.let.rug.nl/~vannoord/alp/Alpino/binary/	OS
36 Persian LG Syntax Parser	fa	Persian Link Grammar Parser is a syntactic parser of Persian. http://students.cs.byu.edu/~jonsafar/	OS
37 Shallow Parser for Hindi	hi	Free Shallow Parser for Hindi released by IIIT-H in Jan. 2008 at IJCNLP. For more details please see resources page is on http://ltrc.iiit.net/ Also check: http://ltrc.iiit.net/showfile.php?filename=release/	F

A.3.4 NERC

Name	Lang	Description	Avail.
38 LingPipe NER	en, ar, zh, nl, de, el, hi, ja, ko, pt, es	NER module of the LingPipe suite. Link: http://alias-i.com/lingpipe	C, F
39 Stanford NER	en	The Stanford Name Entity Recognizer for English. Link: http://nlp.stanford.edu/software/CRF-NER.shtml	F
40 Yooname	en	Yooname is a NER system based on and extending the Balie system. Link: http://www.yooname.com	OU
41 EntityPro	it	EntityPro is a system for the recognition of Italian Named Entities based on Support Vector Machines. Link: http://tcc.fbk.eu/projects/ontotext/entitypro.html	OU
42 SIEMES	pt	Named entity recognizer SIEMES Link: http://poloclup.linguateca.pt/siemes/	F
43 Eihera	eu	Named Entitity Recognizer (NERC) for Basque Link: http://ixa2.si.ehu.es/demo/entitateak.jsp	OU

A.3.5 Semantic Parsers

Name	Lang	Description	Avail.
44 Machinese Semantics	en, fr, de, es, it, nl, sv, da, no, fi	Machinese Semantics is a semantic analyzer. Link: http://www.connexor.eu/	C
45 Shalmaneser	en, de	Shalmaneser (Shallow Semantic Parser) is a supervised learning toolbox for shallow semantic parsing. Link: http://www.coli.uni-saarland.de/projects/salsa/shal/	F
46 WOCADI	en, de	The WOCADI (WOrd ClAss based DIsambiguating) is a syntactic-semantic parser. Demo: http://pi7.fernuni-hagen.de/forschung/wocadi/wocadi_demo.html	?
47 Metamap	en	The Metamap tool can map an arbitrary text to concepts in the UMLS thesaurus Link: http://mmtx.nlm.nih.gov/	F
48 ASSERT	en	The ASSERT system is a automatic statistical semantic role tagger Link: http://cemantix.org/assert	F
49 TeCat	es	TeCat, a toolkit developed for multi-label text categorization (Spanish). Link: http://sinai.ujaen.es/wiki/index.php/Recursos	F

Name	Lang	Description	Avail.
50 WSD Tool	21 EC languages	WSD Tool annotates at the sense level every content word of a parallel corpus in XCES format (RACAI variant). The user can choose to annotate any combination of parts from the parallel corpus for which aligned WordNets exist. Upon request (for research purposes) from RACAI Romania: http://nlp.racai.ro/	F
51 SynWSD	ro, en	SynWSD annotates at sense level all the content words of a given text. It is trained on texts previously annotated with TTL/LexPar and disambiguates texts annotated in the same way. It uses WordNet in XML BalkaNet format as a sense inventory. Upon request (for research purposes) from RACAI Romania: http://nlp.racai.ro/	F

A.3.6 Toolkits

Name	Lang	Description	Avail.
52 GATE	en, de, fr, es, it, zh, ar, ro	GATE is an open source toolkit for IE and text mining; Link: http://gate.ac.uk/	OS
53 STILUS	es, eu, ca, gl, en, fr, pt, de, it, pl, bg, ar, ru	STILUS is a collection of tools for core linguistic processing: parsers, POS taggers, syntactic parsing, disambiguation, text classification, Named Entity extraction... Link (only in Spanish): http://www.daedalus.es/productos/stilus/	C
54 FreeLing	en, es, ca, gl, it	FreeLing is an open source suite of Language Analyzers: tokenization, POS tagging, NER, semantic parsing… Link: http://www.lsi.upc.edu/~nlp/web/ Languages: Spanish, Catalan, Galician, Italian, and English.	F
55 Balie	en, de, fr, es, ro	The Baseline information extraction (Balie) is a multilingual IE toolkit including the following modules: language identification, tokenization, sentence boundary detection, NER. Link: http://balie.sourceforge.net/	OS
56 OpenNLP	en, de, es, th	OpenNLP hosts a variety of NLP tools. Link: http://opennlp.sourceforge.net/	F
57 TextPro	en, it	TextPro is a suite of NLP tools (tokenization, sentence splitting, morphological analysis, POS-tagging, lemmatization, chunking and named entity recognition.) Link: http://textpro.itc.it/	F, C
58 NLCL Tools	en	The NLCL group (University of Sussex) offers a series of tools (morphological processing, parsers) for English. Link: http://www.informatics.susx.ac.uk/research/groups/nlp/resources.php	F
59 SwiRL	en	SwiRL is a Semantic Role Labelling system for English (a suite of syntactico-semantic analyzers: tokenization, POS tagging, chunking, and NERC). Link: http://www.lsi.upc.edu/~nlp/web/	F
60 LingPipe Toolkit	en	LingPipe is a suite of Java libraries for the linguistic analysis of human language (POS tagging, Phrase chunking, NER, etc.). Link: http://alias-i.com/lingpipe	C, F
61 CCG Toolkits	en	The CCG group (Cognitive Computation Group) of the University of Illinois offer different NLP toolkits of interest, including POS taggers, Named Entity taggers and co-reference resolvers. Link: http://l2r.cs.uiuc.edu/~cogcomp/software.php	F

Name	Lang	Description	Avail.
62 C&C Tools	en	The C&C Tools suite consists of the C&C CCG parser (including the computational semantics tool, Boxer) and the C&C taggers (the tools also use the morphological analyser *morpha*). Link: http://svn.ask.it.usyd.edu.au/trac/candc/wiki	F
63 DKPro	de	The Darmstadt Knowledge Processing Repository (DKPro) consists of a number of scalable and flexible NLP components: tokenizer, stemmer, POS-tagger, etc. Link: http://www.ukp.tu-darmstadt.de/software/dkpro/	F
64 ConsILR Pool of Resources	ro	The ConsILR (Consortium for the Romanian Language: Resources & Tools) offers NLP tools for Romanian (lemmatizer,-chunker, etc. Link: http://consilr.info.uaic.ro/	P
65 XCESGen	en, ro	XCESGen, developed at RACAI Romania, is a series of tools to generate parallel corpora in XCES format (metacategories annotation, chunking, lemma/morpho-syntactic label annotation, sense annotation). Upon request (for research purposes) from RACAI Romania: http://nlp.racai.ro/	F
66 MtKit	Depending on availability of parallel XCES corpora	MtKit is an integrated environment that performs lexical annotation/alignment of XCES corpora. It allows the construction of statistical translation models and has an incorporated user-friendly graphical editor. Upon request (for research purposes) from RACAI Romania: http://nlp.racai.ro/	F
67 MTAL	The 21 EC languages	The Multilingual Thesaurus ALigner was developed for EUROVOC English (integral) variant with the incomplete variant for Romanian, allowing the automatic recovery of unique, language independent, CELEX codes for every term; these codes were absent in the Romanian variant of EUROVOC. The system has a general utility in aligning taxonomical conceptualizations with multiple lexicalizations. Upon request (for research purposes) from RACAI Romania: http://nlp.racai.ro/	F
68 TREQ	Depending on availability of parallel XCES corpora	The TRanslation Equivalent Extractor exploits the knowledge embedded in the parallel corpora and produces a set of translation equivalents (a translation lexicon), based on a 1:1 mapping hypothesis. The program uses almost no linguistic knowledge, relying on statistical evidence and some simplifying assumptions. Upon request (for research purposes) from RACAI Romania: http://nlp.racai.ro/	F
69 SearchRoWiki	ro	The web service, originally developed at RACAI for the Romanian shared task at CLEF 2007, searches through the collection of 43000 Romanian documents available on Wikipedia and it is based on a C# port of the Lucene search engine. Link: http://nlp.racai.ro/WebServices/SearchRoWiki.aspx	OU
70 COWAL	---	COWAL is a wrapper of two stand-alone word aligners YAWA and MEBA. COWAL merges the alignments produced by each stand-alone aligner and then uses a trained SVM classifier to prune the unlikely alignment links. The classifier is based on the LIBSVM kit, used with the default parameters. The classifier was trained with positive and negative hand-validated examples of word alignment links. Upon request (for research purposes) from RACAI Romania: http://nlp.racai.ro/	F
71 MEBA	---	MEBA, a C# lexical aligner, is based on an iterative algorithm that uses pre-processing steps: sentence alignment (SAL), tokenization, POS-tagging and lemmatization (through TTL, sentence chunking. Similar to YAWA aligner, MEBA generates the links step by step, beginning with the most probable (anchor links). Upon request (for research purposes) from RACAI Romania: http://nlp.racai.ro/	F

Name	Lang	Description	Avail.
72 LexPar (word linker)	---	PERL application that determines the structure of a connected, acyclic and planar graph of a given sentence, extending the Yuret's algorithm (Lexical Attraction Model). Upon request (for research purposes) from RACAI Romania: http://nlp.racai.ro/	F
73 SAL	---	The sentence aligner is inspired by Moore's program, but it removes its 1:1 alignment restriction, the assumption on the monotonic ordering of the sentences in the two languages, as well as the upper limit on the number of sentence-pairs that can be aligned. It has a comparable precision but a better recall than Moore's aligner. Upon request (for research purposes) from RACAI Romania: http://nlp.racai.ro/	F
74 RACAI web services	en, ro	Linguistic web services for English and Romanian developed at the Research Institute for Artificial Intelligence of the Romanian Academy (RACAI): POS tagging, word linking, WordNet lookup, languages identification, etc. Link: http://nlp.racai.ro/webservices/	OU
75 DIC	ro	Electronic dictionaries compiler created to automatically generate the XML Concede coding from the typographical format (MSWord) of DEX (Romanian Explanatory Dictionary). With minimal transformations, it can be used for the compilation of other dictionaries that use lexicographical and typographical conventions similar to those used in DEX. Upon request (for research purposes) from RACAI Romania: http://nlp.racai.ro/	F
76 PAIL	----	Integrated training environment for Artificial Intelligence, implemented in Common Lisp, containing two modules in the domain of syntactic analysis, based on Augmented Transition Networks (ATN) and chart parsing. The two modules can be combined with the module of automatic theorem prover to create systems for natural language understanding. Upon request (for research purposes) from RACAI Romania: http://nlp.racai.ro/	F
77 Society of Iranian Linguistics	fa	Tools from the Society of Iranian Linguistics: Link: http://www.iranianlinguistics.org/index.cgi	--
78 CLIA Tools	hi, and other Indian lang.	The Cross Language Information Access (CLIA) consortium for Indian languages is producing various tools: Stemmer for Indian languages (Telugu, Hindi), transliteration tool for various Indian languages, etc. (restricted to members of the CLIA consortium.)	P

A.3.7 Languages vs. Tools

The numbers in the table below refer to the identifying numbers of the above mentioned tools.

Lang.	Stemmers, Morph.	POS Taggers	Syntactic Parsers	NERC	Semantic Parsers	Toolkits
en	1, 2, 8	20, 21, 22, 23, 24, 25	28, 29, 30, 31, 32, 33, 34	38, 39, 40	44, 45, 46, 47, 48, 51	52, 53, 54, 55, 56, 57, 58, 59, 60, 61, 62, 65, 74
de	1, 2	20, 21, 23, 24, 26	28, 29	38	44, 45, 46	52, 53, 55, 56, 63
nl	2	21	29, 35	38	44	
fr	1, 2	21	29		44	52, 53, 55
it	1, 2	21	29	41	44	52, 53, 54, 57

Lang.	Stemmers, Morph.	POS Taggers	Syntactic Parsers	NERC	Semantic Parsers	Toolkits
es	1, 2	22	29	38	44, 49	52, 54, 55, 56
pt	1, 2, 3, 4, 5	21		38, 42		53
el		21		38		
da	2		29		44	
sv	1, 2		29		44	
no	2		29		44	
fi	1, 2		29		44	
mt						
pl	1,					53
cs	1, 6	26				
sk						
sl						
bg	1,	21, 27				53
ro	1, 2, 8, 9, 10				51	52, 55, 64, 65, 69, 74, 75
hu	1, 2, 7	26				
et						
lt						
lv						
ru	1, 2	21				53
sq						
ca		22				53, 54
eu	11, 12, 13			43		53
sr						
hr						
tr	2					
ar	1, 14, 15	20	28	38		52, 53
he						
fa	1, 15, 16		36			77
zh	17, 18	20, 21	28	38		52
ja	19			38		
ko				38		
hi[24]			37	38		78
A+[25]	15					56

:

[24] Hindi and other Indian languages.

[25] Other Asian languages: Indonesian, Vietnamese, Mongolian, Kurdish, …

TrebleCLEF Consortium

Istituto di Scienza e Tecnologie dell'Informazione, Consiglio Nazionale delle Ricerche, Pisa, Italy

Università degli Studi di Padova, Italy

The University of Sheffield, UK

UNED, Madrid, Spain

Zürcher Hochschule für Angewandte Wissenschaften, Winterthur, Switzerland

Center for the Evaluation of Language and Communication Technologies, Trento, Italy

Evaluations and Language resources Distribution Agency, Paris, France