ECP-2007-LANG-617001

FLaReNet

D2.1

Up-to-date chart of LR and players and classification along different lines

Deliverable number	*D2.1a*
Dissemination level	*Public*
Delivery date	
Status	*Draft*
Version	*Draft*
Author(s)	*Khalid Choukri, Victoria Arranz, Valérie Mapelli, Hélène Mazo, Djamel Mostefa, Nicolas Moreau*

e*Content*plus

This project is funded under the *e*Content*plus* programme[1],
a multiannual Community programme to make digital content in Europe more accessible, usable and exploitable.

[1] OJ L 79, 24.3.2005, p. 1.

Document evolution

Version	Date	Status	Notes
0.1	02 June 2009	Draft	Internal draft (table of content and synopsis)
0.2	09 June 2009	Draft	Internal version
1.0	16 July 2009	Draft	To be circulated to FlareNet Consortium
1.1	End July 2009	Draft	updated version
2.0	24 September 2009	Draft	Addition of research profiles To circulate to the consortium
3.0	10 October 2009	Draft	Addition of input received from the Steering committee
3.1	13 October 2009	Draft	Addition of input received from various organizations (EAMT, LDC, ISCA)
3.2	15 October 2009	Draft	Final additions with the executive summary
4.0	15 October 2009	Draft	Version to submit to the SC
Revised-01	06 January 2010	Draft	Revised Version with EC comments, submitted to the SC
Revised-10	18 February 2010	Final	formatted version

Table of Content

1 Executive Summary

This document, consisting of two parts, aims at giving the reader a chart of existing Language Resources (data sets and tools, methodologies) and elaborating an overview of the HLT market, mostly from player profile perspectives.

This document follows the two phases of the project with a Phase I that is rather descriptive and aims at providing a critical chart of existing resources by elaborating on the crucial dimensions to describe LRs and Tools and elaborating on the profile (profiles) of HLT key players and area of interest. The phase 2 will exploit these descriptions and work out a model of sustainability analysis to be used for recommendations on potential and future production of LRs, including LRs for emerging areas and new trends.

The picture of the situation and the recommendations that arose from this task served as input to the FlareNet general directive guide for future actions to be considered by the EC (see the FlareNet blueprint for instance).

This document is not an inventory of identified resources though a dedicated section is included. It rather tries to focus on some of the major features that would help understand all issues related to LRs from descriptive metadata to usability in key application, to the composition of various BLARKs for important technologies etc.

After this quick inventory based on some major data centers figures (located in Europe, USA, and Japan, etc.) we elaborate a schema about the different dimensions that help understand the way potential providers could supply such resources and also the way users could express their requirements.

Such schemas elaborate on features such as Data Representation (with the concern of metadata, documentation but also standards and interoperability); it elaborates on various typologies of HLT as being used by a number of major players (data centers, major conferences, major funding agencies, etc.). We try to highlight such aspects through the evolution and trends reported within some of the important conferences, very well representatives of the field.

The BLARK concept is also revisited. In addition to a recall of the concept, we tried to elaborate specific BLARK matrices for some of the Multilingual/Cross-lingual techniques. For the first time, we tried to point out the necessary compromise between the LRs needed by a baseline technology and the performance that one could expect, given the nature, size of the data supplied.

A dedicated section briefly tackles the issues at stake when discussing sharing and distribution of LRs, from various perspectives, with a few words on technical & logistic requirements but a focus on legal, ethical, privacy issues combined with pricing policies.

Although there will be another deliverable on Language Resources self-sustainability, this one introduces the concepts and its main facets. The sustainability model will be discussed in

the other report (due by the end of the project) along lines related to production and sharing models.

The second part of the document is a chart of players and their classifications along various dimensions. We indicate briefly some figures about the market itself (though not the purpose of this report) and then we try to understand the HLT players through their profiles, sectors of interests, location, etc. For this, we exploit input received from many partnering organizations (to which we are very grateful).

2 Introduction

The work of WP2, that is reported on in this deliverable, aims at (1) providing a critical chart of existing resources, classified along various impacting dimensions and (2) elaborating a descriptive schema to describe the profiles of key players and technologies within the European scene.

The sources of input for this document are a thorough analysis and exploitation of various knowledge bases, including brokers' catalogues, existing and available surveys amongst LR producers and users. Among the knowledge used, we relied on ELRA catalogue, ELRA Universal Catalogue, Linguistic Data Consortium Catalogue, National Institute of Information and Communications Technology (NICT) Universal Catalogue and Gengo-Shigen-Kyokai (GSK), OLAC, etc. These sources and the owners will be described in details in the following sections. Identification of Classification criteria/features to categorize LRs along different dimensions will be highlighted through the different practices.

In the terminology used within this report and FLARNET documents we use Language Resources (LRs) to refer to Language data sets, tools of various complexities (in some terminologies/taxonomies we use also terms such as Core versus Integrated technologies to define such complexity). We may even see this definition of resources extended to guidelines, standards, processes but these are not part of this document. We feel that the Human Language Technology (HLT) field evolved a lot and thus it is not surprising that the terminology did follow/anticipate such changes. As an example of such list, please find herein the one derived from the analysis of LREC submissions over the last few editions, these are examples and do not pretend to be an exhaustive list to describe all existing Language Resources:

- Corpus (single modality, multimodalities, etc.)
- Lexicon
- Terminology
- Ontology
- Grammar/Language Model
- Annotation Tool
- Speaker recogniser
- Signal Processing/Feature Extraction
- Transcriber
- Image Analyser
- Representation-Annotation Formalism/Guidelines
- Representation-Annotation Standard, Best Practice

- Tokenizer
- Tagger/Parser
- Named Entity Recogniser
- Word Sense Disambiguator
- Language Identifier
- Prosodic Analyser
- Metadata
- Evaluation Data
- Evaluation Tool
- Evaluation Package
- Evaluation Methodology, Formalism, Guidelines
- Evaluation Standard/Best Practice

Although a first very short section indicates some statistics about existing LRs, this report is not an inventory of resources.

We will point out the content of catalogues of the two major data centers ELRA (http://catalog.elra.info/), LDC (http://www.ldc.upenn.edu/Catalog/) and then of NICT (http://www.nict.go.jp/), that has started a project similar to the ELRA universal catalogue and in partnership with ELRA and LDC. The information about the tools (as for instance catalogued by the DFKI project with the LT-World (http://www.lt-world.org/) will be incorporated in a coming version if need be.

This report consists of two chapters. The first one elaborates on the description of various aspects of LRs and aims at identifying those areas and resources that are of paramount importance to the HLT field. In a coming report the issue of self-sustainability will be tackled and a "model" of sustainability analysis will be sketched to be used for the future production of LRs with clear factors and features impacting on sustainability.

An important part of the current report relates to the BLARK concept that addresses issues discussed during the Vienna FlareNet event, such as the types, quantities, quality of resources that constitute a BLARK for a given language. The costs of a BLaRK for a given language and a set of technologies are also taken into consideration. According to our current knowledge of state of affairs, we will elaborate on our perception of the language resource landscape listing for some technologies the needs and requirements.

The outcome of this first chapter will be an updated chart of key LRs available both at EU and non-EU languages, with a classification of existing LRs in terms of type, medium, languages, application(s). This will imply a later analysis and revision of the suitability of issues such as available metadata sets (ELRA, CLARIN, OLAC, NICT, and possibly INTERA), data storage formats, and character encoding systems.

The second chapter aims at describing the player profiles, structures, key areas of involvement, etc. It mainly focus on "pure" technology (Automatic Speech Recognition, Machine Translation, Information Retrieval, etc.) but may include players that integrate such technologies within areas like voice applications (transcription of news, parliamentary speech, interactive dialogues and call centers, multimedia and online publishing (content producers), translation services and products, etc.

The main goal of this chapter is to draw a clear and accurate (even if not exhaustive) picture of the main players and actual as well as potential users, either private or public, involved in the language industry. The chapter also aims to position the European players vis-à-vis their competitors from the US and Japan.

The main focus of this chapter is derived from our knowledge of the LRs market(s), including market surveys that were conducted regularly over the last decade. The important facts collected so far consist in the inventories, the needs and requirements, the trends, and overall the size of traded resources versus the non traded ones. To extend this beyond the ELRA & FlareNet circles, other organizations have been approached and their contributions are used to draw the summarized picture given in this report.

3 Chart of LR and classification along different lines

3.1 Some quick facts & figures about several Language Resources

The idea behind these quick facts is to show how the HLT domain structured itself under the incentives of data centers that initially (and primarily) collected information about LRs and later on catalogued them. This also showed the need for harmonized metadata that is still a crucial and open issue. The catalogues do show how things evolved over the last ten to fifteen years.

3.1.1 ELRA catalogue

Since October 1996, ELRA managed to sign the following number of agreements with providers:

Agreements	Speech	Written	Terminology	Evaluation Packages
30th September 2009	163(1)	86(2)	12	15

The number of resources in the catalogue is given below:

Resources	Speech	Written	Terminology	Evaluation
30th September 2009	446(1)	276(2)	290	36

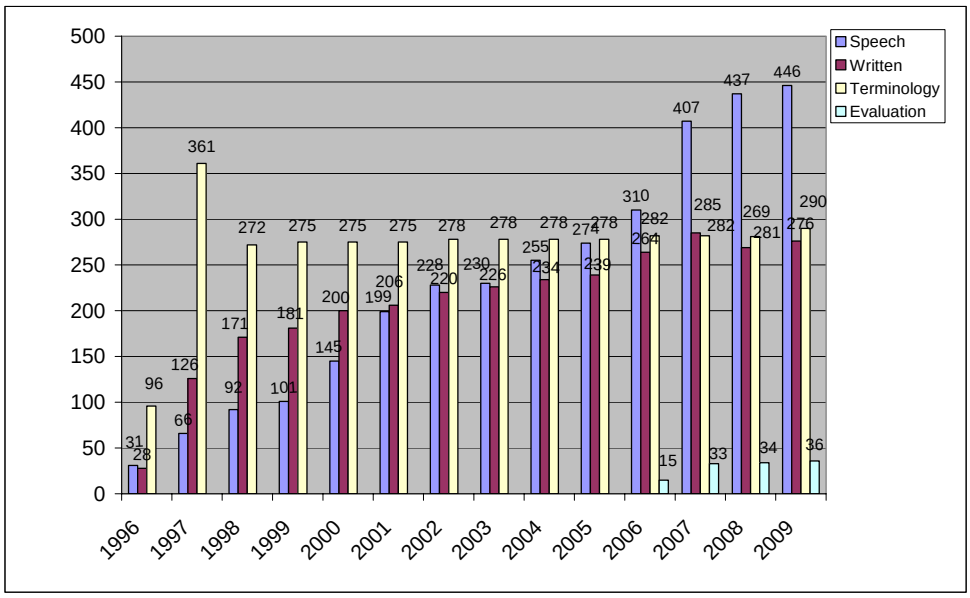
(1) Including Multimodal-Speech oriented databases.
(2) Including Multimodal-Written oriented databases.

The written resources consist of the following items:

Written Corpora	Monolingual lexicons	Multilingual lexicons
62	81	133

.

The increase of the number of resources over the years is illustrated below with the four major categories: speech, written, terminology and evaluation packages.



3.1.2 The ELRA Universal Catalogue

The ELRA Universal Catalogue (referred to herein as Universal catalogue or UC) comprises information regarding Language Resources identified by the ELRA team. It aims to be a repository for all identified Language Resources so as to provide ELRA members (and beyond) with information on what resources exist and what their features are.

All throughout 2008, efforts have been increased on the identification of Language Resources and the setting up of a public catalogue that will serve the whole HLT community in its search for Language Resources. This long-term task requires constant work and coordination. For 2008, the plan was to focus our attention on identifying a minimum of 30 LRs per month and ensuring that the Universal Catalogue is updated with all information related to these identified resources. This objective has been achieved having increased the number of new resources in our catalogue by 256 resources in just 7 months (about 36 per month). Intense identification efforts were mostly devoted to the interaction with its external users, while still remaining alert to identification tasks.

As of end December 2008, the Universal Catalogue comprises 1,307 identified resources. The distribution of these figures according to the type of resources is given in the table below:

LR Type	# of LRs
Tools	58
Speech	425
Written	776
Terminology	23
Multimodal/Multimedia	25

The Universal Catalogue was opened to the general public on the 1st October 2008. To this date, this catalogue had been a service exclusively offered to the ELRA Members. Since October, the whole Language Resource and HLT communities can benefit from this service, which aims at helping all researchers, developers or merely interested users from both communities to gain access to the existing resources in the world and the information related to them.

3.1.3 The Linguistic Data Consortium Catalogue

LDC catalogue contains over 445 (public) resources (up to October 2009) that are classified along similar categories as given above with some additional features. The list of such categories is:

- lexicon
- lexicon, speech, text
- speech
- speech and text
- speech and transcripts
- speech, text
- text
- transcripts
- video
- video, text

As we can see, each category is distributed over a number of modalities and/or combinations of modalities. The following sub-categories are even more explicit on the activities carried out by LDC members:

broadcast conversation
broadcast conversation speech
broadcast conversation text
broadcast news
broadcast news speech
broadcast news speech, text
broadcast news text
broadcast news transcripts
broadcast news video, text
dictionaries
dictionaries lexicon
dictionaries text
email
email text
field recordings
field recordings lexicon
field recordings speech
field recordings video
government documents
journal articles

journal articles text
lexicon
meeting speech
microphone conversation
microphone speech
news magazine
newsgroups
newswire
telephone conversations
telephone speech
transcribed speech
transcribed speech
varied
video
web collection
weblogs

3.1.4 The Japanese multiple efforts

3.1.4.1 The National Institute of Information and Communications Technology (NICT) Universal Catalogue

NICT was newly launched as the National Institute of Information and Communications Technology (NICT), an incorporated administrative agency.

NICT was established by the Japanese government "*to carry out research and development in the field of information and communications technology, which supports the upcoming ubiquitous network society in an integrated manner from basis to application and also provides comprehensive assistance to the public and private organizations working in this field*". NICT newly started the 5-year medium-term plan in April 2006. In this important turning point, NICT integrated the contents of research and development already performed up to now into 3 research domains such as "New Generation Network Architecture Technology", "Universal Communications Basic Technology" and "ICT for Safety and Security" and reviewed and very much improved the research organizations to promote these research domains. From our perspective, we should know that NICT somehow inherited activities in the human language technology area that were trusted in the past to the famous private corporation ATR (Advanced Telecommunications Research Institute International).

From the Language Resource perspective and from the NICT UC descriptions (http://facet.shachi.org/?ln=en) we understand that the information collected by NICT (and referred to as "Shashi") consists mostly in harvested data (a la OLAC). So far we can see over 2700 resources listed, most of them coming from ELRA, LDC, GSK, and others. NICT is offering fair links to the data centers from which such resources can be licensed.

3.1.4.2 Gengo-Shigen-Kyokai (GSK)

Gengo-Shigen-Kyokai (GSK) (literally "Language Resources Association") was established in June of 2003. By promoting the distribution of language resources such as speech data, lexicons, text corpora, terminology, and various tools for language processing, GSK aims at contributing to speech / natural language processing technology, research, and industrial development that require those language resources, as well as to the advancement of research in the linguistic field.

Although the official catalogue of GSK is restricted to a few corpora and lexica, GSK is a key player in Japan and can mediate for the acquisition of a large number of the resources that are owned by different Japanese institutions and agencies.

3.1.5 OLAC

In addition to these major players, we can also mention the work being done by OLAC (the Open Language Archives Community), *"an international partnership of institutions and individuals who are creating a worldwide virtual library of language resources by: (i) developing consensus on best current practice for the digital archiving of language resources, and (ii) developing a network of interoperating repositories and services for housing and accessing such resources."*

OLAC is not another catalogue but rather a coordination body. A number of data centers (e.g. ELRA, LDC; ELRA and LDC are represented on the Advisory board of OLAC) agreed to join forces and allow the users to access their data through the OLAC portal. ELRA and LDC catalogues are exported daily onto the OLAC metadata schema and made available to all interested parties. Many other archiving houses joined and allow their users to access information through the OLAC portal. Most of these archives are involved in field linguistics, social sciences, and humanities.

For the languages covered by OLAC, we have more than 20,000 items (out of which less than 3000 are available online and even less suitable for HLT purposes). We can see more details on the following table that gives the coverage in relation to language size and some "extinct" languages (of importance to Social sciences and Humanities). Of course, most of these resources are suitable only to social sciences and humanities.

Population range	*Languages*	*In OLAC*		*Items*
10,000,000 or more	83	82	99%	3,341
1,000,000 to 9,999,999	264	223	84%	1,431
100,000 to 999,999	892	575	64%	2,607
1,000 to 99,999	3,746	1,797	48%	9,012
100 to 999	1,071	392	37%	2,305
1 to 99	548	271	49%	832
Unknown	308	86	28%	307
All living languages	*6,912*	*3,426*	*50%*	*19,835*
Extinct languages	602	130	22%	315

3.1.6 Other initiatives

CLARIN has also started a competing activity (both with the HLT data centers and OLAC). More details could be provided in a later version of this report depending on the CLARIN work progress.

In China, the Chinese-LDC plays a role that is similar to ELRA and LDC, while focusing on the supply of resources for technology evaluations conducted within national programs.

For Korea, the Speech Information Technology & Industry Promotion Center (SiTEC) was founded at Wonkwang University in 2001 with the aid of the Ministry of Commerce, Industry

and Energy to promote speech information technologies and support industries. SiTEC is involved in the creation and distribution of speech corpora.

Other initiatives exist in India, Thailand, Indonesia, South-Africa, and many other countries.

Last but not least, some very important geopolitical/geographical areas (but not necessarily economically with high GDPs) do not seem to be well equipped with LR centers such as South America and the Arabic region.

3.1.7 Conclusions and Perspectives

From these "surveys", we have seen the huge variety of LRs and the related technological areas but also the large number of descriptive schemas, including the various metadata elements. One of the crucial and expected contributions of FlareNet is to support projects on harmonization of such issues that is being conducted by major players like ELRA & LDC. There is a room for strong initiatives in conjunction with FlareNet and the potential Open Resource Infrastructure that is being promoted (a major task within the possible T4ME network). Such contribution will be more than useful to allow researcher have access to resources without mastering multiple metadata sets and proliferation of search engines. Although our main concern is LRs for HLT, it is important to consider the achievements of OLAC with respect to a minimal coordination of international initiatives.
It is also important to point out that the descriptions reported herein are inspired from the traded resources and one can imagine the complexity of the context for non traded resources. [1]

Part of this work will be done for FlareNet during the second phase at the already planned events: Brandeis workshop (11/2009) with ELRA & LDC focusing on metadata and related matters and LREC'2010 (05/2010) with the so called LREC'Map[2] to gather more input about LRs being produced, used, re-used, customized, etc. by the LREC participants (over 1000).

3.2 LRs Dimensions

A small but crucial part of the metadata issues introduced above concerns some of the features that are often neglected (e.g. documentation for users and machine-readable). This section introduces some of these features and, again, illustrates, the variety of approaches used by the major players for this purpose.

[1] In a Internal Market analysis conducted for ELRA by Bain in 2002, the ratio between publicly traded (and visible) and non-traded resources is 30%-70%, which showed the market potential growth.

[2] The LREC 2010'Map is intended to monitor the use and creation of language resources (datasets, tools, etc.) and is based on the LREC submissions that require authors to submit information on both existing resources that are used and that are created.

The idea behind this is to collect as many dimensions and typology characteristics as possible to design and boost a harmonization initiative within the strong cooperation between ELRA and LDC, with the support of FlareNet and the partnering US project SIL [1].

It is also essential to monitor the trends of the HLT various areas (as will be illustrated in particular by section 3.2.4) thanks to the strong interest of major scientific organizations with which we liaise already on this issue e.g. ACL, ISCA, EAMT, and exploiting the useful data they collect through their own means (members, conferences, Special Interest Groups, etc.).

3.2.1 General introduction & data representation

In order to represent and ensure some interoperability and sharing of LRs between different players, it is essential that the data be represented in a common way. The data is usually understood as a set of constituents: the "atomic units"[2] that constitute the LRs, the information that documents such LRs from all perspectives and possible/potential views, the media that stores the data using a given format, etc. Some of these dimensions are more critical than others for a given category of users. In the best context all these dimensions are documented, in the worst none of these exists (except the data itself and the storing media) and the re-use of the data by others than the data owner/producer requires a huge effort of reverse-engineering to understand the data format and the data content.

It is important that commonly accepted practices are used for these purposes and hopefully some may become standards (or best practices) within the community. It is according to this perspective that ELRA has been promoting its "quick quality check" (QQC) process that offers a basic quality assessment of a LR that could be sufficient for a potential user to obtain an overall impression of the quality of the resource through a descriptive approach to the dimensions mentioned above. This QQC is a procedure that could be used by producers (and later on checked by independent bodies) to ensure that the minimal critical dimensions are documented, such as availability/reliability of information on technical, formal issues such as media, number of files, file structure(s), file names etc. Depending on the nature of LRs, more factual information should be documented and checked.

As an example, ELRA recommends to ensure that a clear and suitable documentation be prepared such as reports on (at least) critical aspects of the metadata (see below the section on metadata), owner/copyright holder, format and encoding issues of the data and the files, languages(s) covered, domains, intended applications, applications in which the data was used, etc. The documentation (and the QQC reports) should also report on the formal parameters that have to be verified (e.g. for a lexicon the list of legal values and attributes, the obligatory fields, etc.). The documentation should also reflect the "reliability" of the "linguistic" information e.g. any annotation should be documented from both the formal and the content aspects.

---

[1] "Sustainable Interoperability for Language Technology" (SILT) is funded by the National Science Foundation under the Community-based Interoperability Networks (INTEROP) program:

(www.nsf.gov/pubs/2007/nsf07565/nsf07565.pdf).

[2] The LRs could be a list of lemmas/words for a lexicon, a set of words or sentences for a textual corpus, a set of audio recordings or samples of speech in a speech database etc. The atomic units are then a lemma, a word, a sentence, a speech waveform or a sample of a speech waveform, etc.

In order to contribute to the dissemination of such good practices, ELRA has promoted the use of an extensive metadata set for the language resources it catalogues, the performance of quick quality checks to ensure that such metadata aspects are consistently used and that the metadata parameters are comprehensively filled in including the general documentation (or the manual) of the LRs.

The dimensions we consider as important to chart the LRs are:

(1) typologies of resources and the related technologies/applications (used for, potentially usable in, etc),
(2) metadata and descriptive documentations,
(3) coding/encoding schemas and formats,
(4) languages,
(5) domains of coverage, etc.

Given the spirit of the FlareNet project, these items could be revised and augmented with suggestions from other contributors once the draft is made public through our wiki service.

3.2.2 Typologies of Human Language Technologies

There are a number of crucial and important points that should be listed for each LR:

- all the applications for which the data may have been specified and designed.
- all the applications in which it has been used;
- all potential applications for which it could usefully be exploited.

For instance, a set of multilingual broadcast news recordings (as one can get from Voice of America or Radio France International) can be used to train or evaluate Automatic Speech Transcription technologies (speech to text) but could also be used to assess performance of language identification systems.

A non-exhaustive list of possible applications that can encapsulate and incorporate Language Resources is given below. This list is derived from the one used within the ELRA catalogue.

- Automatic person recognition
- Automatic speech recognition
- Avatar synthesis
- Constituent Recognition
- Diacritizer
- Dialect / language identification
- Dictation
- Discourse analysis
- Embedded speech recognition
- Emotion Identification
- Expression recognition
- Face recognition
- Face verification
- Grapheme-Phoneme Converter
- Humanoid agent synthesis
- Information retrieval
- Language Aligner
- Language identification
- Lip tracking analysis
- Lips movement
- Machine Translation
- Morphemes Disambiguator
- Morphological Analyzer
- Multimedia development
- Parsers and Grammars
- Phonetic Grammar
- POS tagging
- Pragmatic analysis
- Proper Names Recognition
- Prosodic Phrasing
- Referent Resolution
- Semantic Analysis
- Sentence Boundary Detection
- Shallow parsing
- Speaker Adaptation
- Speaker audio/visual identification, verification,
- Speaker audio/visual tracking,
- Speaker turn detection
- Speaker audio/visual verification
- Speech assisted video control
- Speech enhancement
- Speech indexation
- Speech Input
- Speech / Nonspeech detection
- Speech recognition
- Speech synthesis
- Speech understanding
- Speech verification
- Speech/lips correlation analysis
- Spoken dialogue systems
- Talking head synthesis
- Telephony speech applications
- Text generation
- topic detection, segmentation
- Transcription of broadcast News
- Transcription of conversational speech
- Voice control
- Word Meaning Disambiguation
- Word spotting / boundary identification

While working on the LREC'2010 edition (May 2010, Malta, www.lrec-conf.org) and on this quick list of applications, further topics, keywords, related to the use of LRs in technologies have emerged as the outcome of a quick brainstorming. These are given herein for information purposes.

- Acquisition
- Dialogue
- Discourse annotation, representation and processing
- Document Classification, Text categorisation
- Emotion Recognition/Generation
- Information Extraction, Information Retrieval
- Knowledge Discovery/Representation
- Language Identification

- Language Modelling
- Machine Translation, SpeechToSpeech Translation
- Multimedia Document Processing
- Named Entity Recognition
- Natural Language Generation
- Parsing
- Person Identification
- Question Answering
- Semantic Web
- Sign Language Recognition/Generation
- Speech Recognition/Understanding
- Speech Synthesis
- Summarisation
- Text Mining
- Textual Entailment and Paraphrasing
- Topic Detection and Tracking
- Voice Command and Control
- Web Services
- Word Sense Disambiguation
- Other (specify here)
- Not Applicable

If we review the LDC list of applications, we definitely identify a large overlap but also some distinctions, many of them having to do with the projects that LDC carried out for the various USA agencies:

- anaphora resolution
- automatic content extraction
- content-based retrieval from digital video
- cross-lingual information retrieval
- discourse analysis
- discourse parsing
- distillation
- finite state technology
- gesture recognition
- gesture synthesis
- information detection
- information extraction
- information extraction from video
- information retrieval
- instruction
- language generation
- language identification
- language modeling
- language teaching
- linguistic analysis
- machine learning
- machine translation
- meeting summarization
- message understanding
- metadata extraction
- morphology
- morphology learning
- named entity recognition
- natural language processing
- nominal expression generation
- parsing
- part of speech tagging
- phonetics
- phonology
- pragmatics
- pronunciation modeling
- prosody
- psycholinguistics
- question-answering
- sociolinguistics
- speaker identification
- speaker segmentation and tracking
- speaker verification
- speech recognition
- speech synthesis
- spoken dialogue modeling
- spoken dialogue systems
- standards
- subjectivity analysis
- summarization
- syntactic parsing
- tagging
- temporal parsing
- temporal reasoning
- topic detection and tracking

From the ACL perspective, we can also list the topics disseminated within their "calls for papers" and for which they got high-level scientific publications. ACL puts further emphasis on the written technologies, though speech, gesture, sign languages and multimodal processing are indicated as well:

Black-box evaluation of systems in application settings
Corpus-based language modeling
Corpus-based parsers and evaluation
Cross-language information retrieval
Development of language resources,
Dialogue systems for collaboration, tutoring and behavioral
Discourse and Pragmatics
Discourse, dialogue, and pragmatics
Embodied conversational agents, virtual humans and human-robot conversation
Evaluation methods and user studies
formal semantics & logic
Formalisms and Metrics
General information retrieval
Glass-box evaluation of systems and system components
Grammar engineering
Grammar induction and development
Information extraction
Information retrieval
Information retrieval/NLP applications
Intelligent systems for natural language interaction, including intervention
Knowledge acquisition
Language Generation
Language modeling for spoken language
Language modeling for text processing
Language processing in domains such as bioinformatics, legal, medical, etc.
Language resources, evaluation methods and metrics, science of annotation
Language-enhanced platforms for interactive narrative and digital entertainment
Large scale language processing
Lexical and knowledge acquisition
Lexical and ontological semantics
lexical semantics
Lexical/ontological/formal semantics
Machine translation
Machine Translation and Multilingual processing, including
Machine translation of speech and text
Mathematical linguistics and grammatical formalisms
Mining from textual and spoken language data
Multilingual language processing
Multi-lingual speech recognition and language identification
Multimodal language processing (including speech, gestures, and other communication media)
Multimodal representations and processing, including speech and gesture
NLP applications and systems
NLP in vertical domains, such as biomedical, chemical and legal text
NLP on noisy unstructured text, such as email, blogs, and SMS
NLP-oriented information retrieval
Parsing algorithms and implementations
Phonology/morphology, tagging and chunking, and word segmentation
Psycholinguistics
Question Answering
Rich transcription (automatic annotation of information structure and sources in speech)
Rich transcription and spoken information retrieval
Science of annotation
Semantic role labeling
Sentiment analysis and opinion mining
Sentiment/attribution/genre analysis
Speech generation and synthesis
Speech processing
Speech recognition

Speech translation
Speech/MT-oriented information retrieval
Spoken language processing
Spoken language understanding and generation
Statistical and machine learning methods
Statistical and machine learning techniques for language processing
Summarization
Syntax, parsing, grammar induction
Text Data Mining, Information Extraction, Filtering, Recommendation
Text mining and natural language processing applications
textual entailment & paraphrasing
Textual entailment and paraphrasing
Topic/text classification and clustering
Treebank and corpus development
Treebanks, proposition banks, and frame banks
User studies
word sense disambiguation

The idea behind these long lists of topics, activities, and areas of research is to draw the reader's attention to how things are structured within the HLT field. Tasks to derive a reliable taxonomy through a clustering of these individual themes are conducted by different labs. Hopefully, FlareNet could play a federating role to promote one version that could obtain the consensus of the major centers (without referring to it as "standardized" version).

In addition to the data sets mentioned herein, one should keep in mind that related to those applications, a number of other specific items have to be taken into consideration, such as data for training and data for evaluation, language(s), domain of activities, topics, genre, time period that the data cover, etc.

3.2.3 Typologies of LRs for Human Language Technologies

As quickly introduced above, LDC and ELRA, as the major players acting as data centers, have organized their catalogues historically by classifying the resources according to their main "modality".

Such categorization was established as follows:

- Speech-oriented LRs: these are either recorded data, most of the time coming with their written transcriptions, as well as pronunciation lexica. Different recording media have been identified under this category, such as telephone (fixed/gsm/umts), microphone of various qualities, broadcasting, etc. The crucial issue here is the sampling frequency, number of bits of quantification, the acoustic conditions, the speakers' profiles, etc.

- Written-oriented LRs: In this area, we count more specifically written corpora and written lexica. More and more, terminology data are also gathered under this category and referred to as specialized lexica.

- Multimodal/Multimedia resources: these are recordings and annotations of a combination of several modalities such as speech/audio, Video/image, hand gesture, facial expression, body posture, etc. Those different modalities are

acquired through different types of media, which adds a multimedia aspect to that category (the modalities can be recorded through different media such as microphone, camera, video-recorder, specialized devices for handwriting, etc..). During the last couple of years (and noticed through the number of LREC satellite workshops/special sessions) we have seen the emergence of sign languages that are in between multimedia resources (images) and lexica (association of a "lemma/word" and information).

The evolution of these areas and the importance of each of them with respect to the research activities being conducted is highlighted by the scientific productions, in particular by the papers published in the major conferences (in addition to the categories of resources acquired by the scientific communities). While some areas have started their own specialized satellite events (Multimodal issues are addressed within the ICMI-MLMI: multimodal interaction, systems and methods, Interspeech organizes special sessions related to speech resources, sign languages have started dedicated events, etc.), ELRA remains the major event on LRs and Evaluation of technologies ensuring the reliability of the collected facts.

3.2.4 Evolution of the HLT arena scientific productions & scientific trends

Let us also highlight these trends and evolutions through the publications submitted to some of the major conferences in the areas e.g. LREC, InterSpeech, and EAMT

Looking at the last two issues of REC conferences[1], organized by ELRA, a number of main areas were highlighted where Language Resources are a major component. The table below shows the evolution of number of submissions with respect to the different types of resources and fields where resources are considered as a main component:

[1] http://www.lrec-conf.org/

Submissions 2006			Submissions 2008		
Track	**Number**	**%**	**Track**	**Number**	**%**
E	165	21,02	**E**	180	20,00
G	42	5,35	**G**	163	18,11
MM	38	4,84	**MM**	49	5,44
S	101	12,87	**S**	114	12,67
T	69	8,79	**T**	72	8,00
W	370	47,13	**W**	322	35,78
Total	785	100,00	**Total**	900	100,00

Legend:

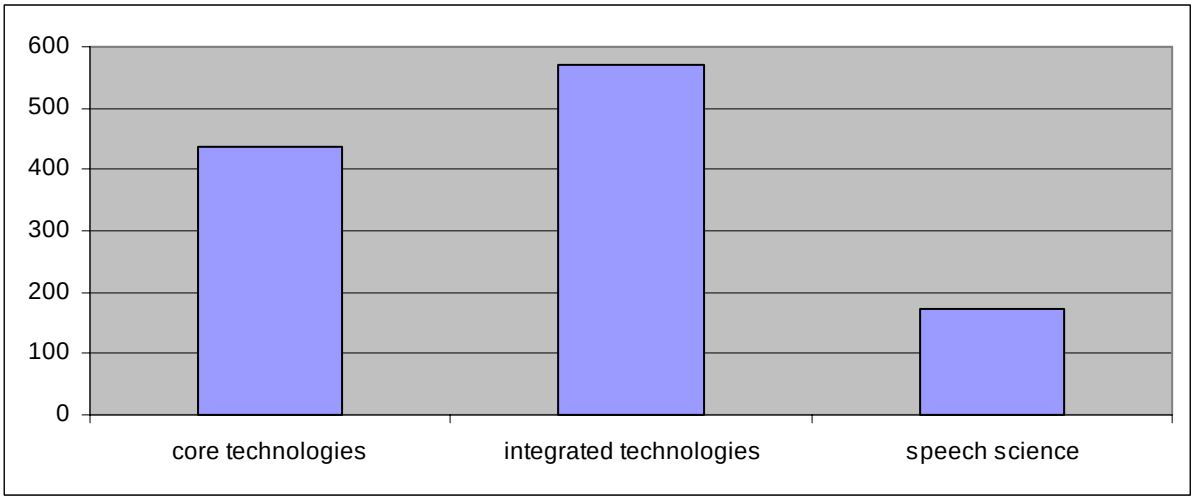
E = Evaluation
G = General issues (project, infrastructures, etc.)
MM = Multimodal/Multimedia resources
S = Speech resources
T = Terminology resources
W = Written resources

This table shows some stable and high participation of organizations in the fields of written resources, as well as evaluation activities. However, the interest remains high and constant for Speech resources.

The European Association for Machine Translation (EAMT)[1] continues to organize very successful conferences in addition to the general MT Summit. Over the last few years, we have seen a strong shift to Machine translation based on pure statistical modeling of aligned texts in different pairs of languages. From the recent EAMT (and associated) summits, we see there is a clear move-away from pure phrase-based MT to more hierarchical models, where syntax is incorporated by other means (in the source language). At the last 2009 Summit, there was a clear indication from all users that MT was proving useful to them. The only difference was the extent of their savings (from 10% to 150%, although one should be cautious about the meaning of these figures). Furthermore, there is a clear evidence of research groups (Edinburgh, Canadian National Research Council (Gatineau), Dublin City University) working closely with industrial partners to make their good research-based systems of real use for end-users. Evaluation is also moving away from the "string-based metrics" that have been in use to more complex ones. A huge effort is devoted to this topic within the Euromatrix+ project (http://www.euromatrixplus.eu/), a project supported by the EU Information Society Technology program under FP7.

[1] the authors are very grateful to Andy Way (EAMT president) for his input on this issue.

If we consider the major International conference on speech sciences & technologies (Interspeech), we can see that the 2008 event attracted over 1200 papers distributed as follows:



Speech science represents the basic R&D activities and comprises topics like: Human speech production, Speech perception, Phonology, Phonetics, Cross-language comparisons, Physiology and pathology, Spoken language acquisition, development and learning, Discourse and dialogue, Prosody, Paralinguistic and nonlinguistic cues, Speech and other modalities (e.g. gestures, facial expressions).

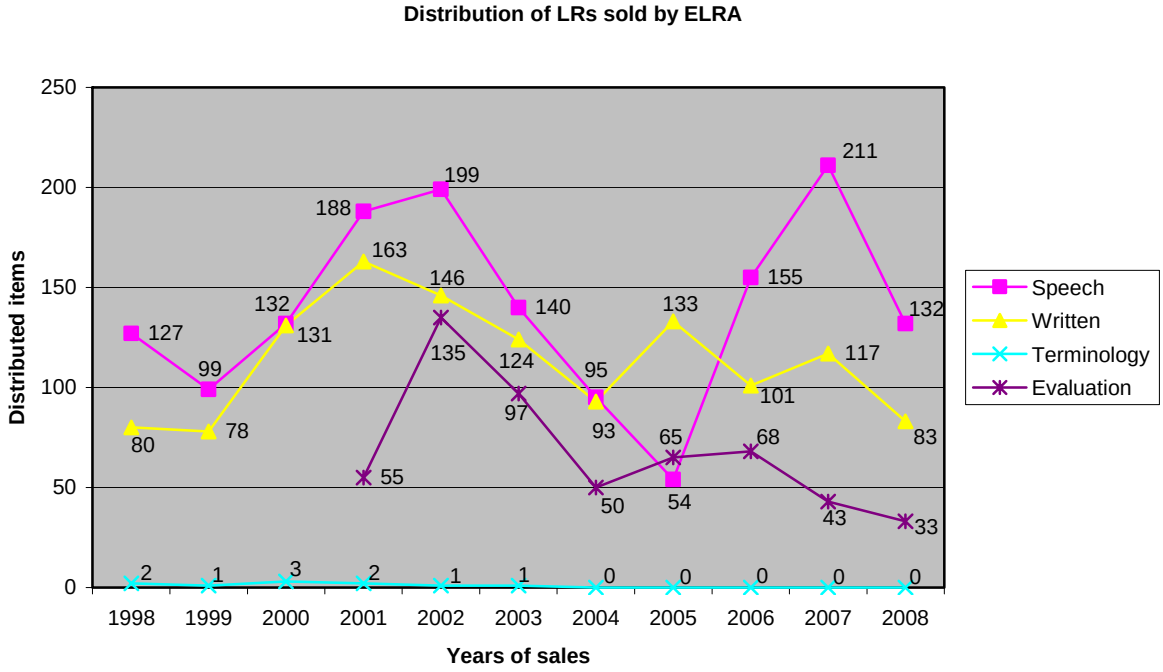
Core technologies comprise the development of basic tools as in Speech & audio analysis, Multimodal and multimedia signal processing, Speech enhancement, Speech coding & transmission, Accent & language identification, Speakers: expression, emotion & personality recognition, speaker verification & identification, Speaker voice conversion and modification, SL generation, Automatic Speech Recognizers.

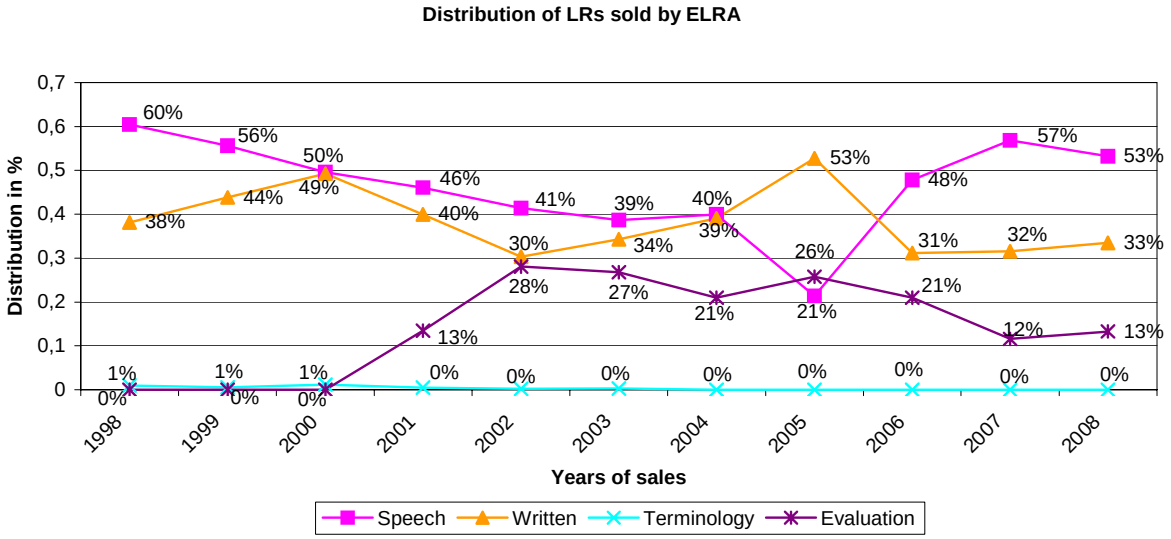
Integrated technologies comprise the development (and even the deployment) of systems, applications, such as Spoken dialogue systems, Systems for Large Vocabulary Continuous Speech Recognition (LVCSR) and rich transcription, Systems for language information retrieval, Systems for Spoken Language understanding & summarization, Systems for spoken language translation, Applications for aged and handicapped persons, Applications in education and learning, Applications in other areas.

Moreover, within this “applications” category, ISCA inserted papers about Speech resources & annotation as well as Evaluation & standardization of Spoken Language technology and systems and reported about 35 publications (24 & 11 respectively) on these two topics.

It is also important to point out the total number of papers on the technology theme (core & integrated) which was over 1000 despite the strong consolidation of the speech technology sector that seems to have impressively shrunk in terms of players over the last few years.

Let us look at this aspect from another perspective: acquisition and/or distribution of Language Resources. If we consider the distribution data from ELRA, we can also get a picture of the way LR types are acquired:



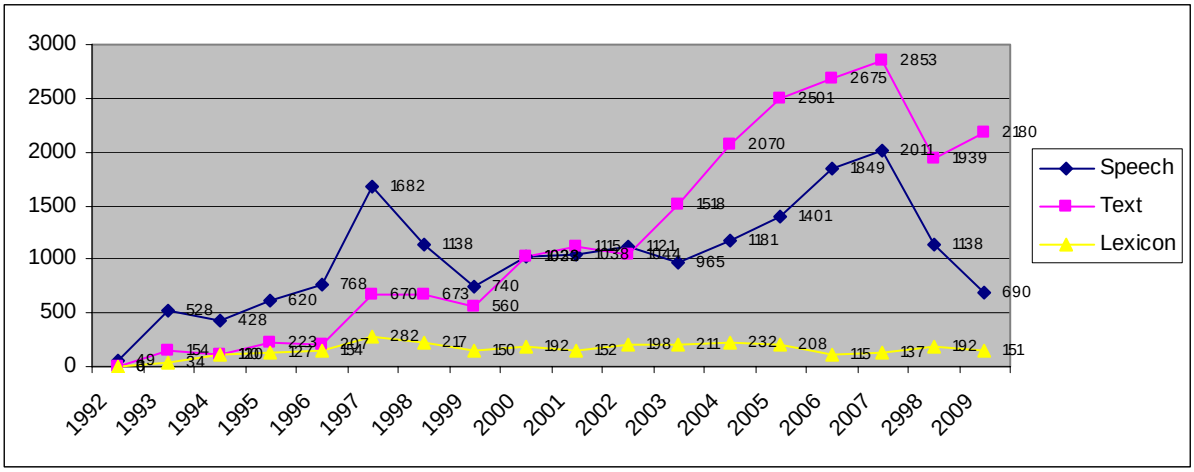


When discussing and comparing the distribution issues, in particular of ELRA and LDC, one should bear in mind that the two organizations have different membership policies and different distribution policies: ELRA's membership does not entitle members to get resources free of charge but rather get them an important discount on the public prices. LDC members get de facto a number of resources free of charge. LDC has elaborated a number of

membership scenarios with different profiles and status. For instance, Standard Members are entitled to receive one copy of each corpus released in their years of membership with a maximum allowance of 16 corpora per membership year. There may be additional charges for corpora owned or produced by others and distributed by LDC. Additional corpora beyond the 16 corpora may be licensed at the individual corpus licensing fee; Subscription Members automatically receive two (2) copies of each corpus released in their years of membership, etc. etc.)

Given these facts, if we consider the overall distribution of LDC since 1992, we have the statistics illustrated by the diagram given herein. The LDC corpus distribution contains a breakdown by year (1992- October 10th, 2009) of the number of corpora LDC distributed. The data appears by corpus type (speech, text (written), lexicon). The corpora reflected in the sheet are limited to those in the LDC public catalog. LDC also distributes corpora for evaluations, projects, etc. LDC distributed over 42k copies of releases from the public catalog during the period and this is approaching 65k copies distributed overall when evaluations, projects and other releases are included.
N.B. ELRA and LDC use different labels, ELRA's data that compares with LDC's is Written versus text + lexica.

In terms of determining trends for speech v. text, etc, there may not necessarily be a correlation between LDC distribution and user demands. LDC aims to release both types of data every year, but this is not always possible. Therefore, the fact that more text corpora were distributed in a given year may reflect the fact that LDC releases that year tended to be text data sets.



3.2.5 Conclusions and Perspectives

To sum up, we can see from these concise facts and figures how the research (at least as seen from the major scientific events, combined with the activities of data centers) have evolved over time. A clear conclusion from the LREC data (very representative of the HLT field) we can point out the importance of evaluations conducted within the community, both

through well designed and formalized campaigns and internal “quick” evaluation to assess progress. Nevertheless, more input required in particular on resources such as Multimodal/multimedia (Video annotations, meeting minutes, etc.), Sign languages, SMS type of language, OCR & Handwritten modalities, etc. In addition to this, we will be working, during phase 2, on these issues from the language perspective as we strongly see the lack of essential items for almost all languages. The LREC’map mentioned in previous sections will help us monitor this over time with a snapshot at least every two years (other organization are encouraged to adopt such procedure so such snapshots would be available over other dimensions).

3.2.6 The Metadata and Coding Dimensions

To enable a better interoperability of language resources (better sharing of information and storing of language resources), a number of standardized description features need to be defined, associated to access criteria and rights to such features. Such standardized features are known as “metadata sets” and consist of “data about other data”. These are commonly used for cataloguing language resources and enabling information retrieval through various tools. Such metadata are strongly connected to the domain of activity of those who are potential users so they can easily retrieve the data through such descriptors and hopefully share such concern about data description with the producers/providers. In some cases producers/providers of LRs are not interested in adding detailed metadata features to their resources as this is only required when the data is shared across individuals, organizations, etc. Moreover, data distribution centers sometimes have to take in charge such activity of adding metadata descriptors to the catalogued resources.

Recently, part of this work has been brought for discussion within ISO. ELRA (also as part of FlareNet) is contributing to these efforts through its liaison expert Nicoletta Calzolari.

The metadata use distinguishes various levels of granular descriptions; at least two of them are of paramount importance in our work within FlareNet: (a) metadata related to language resources (e.g. description of a lexicon, a database of speech recordings, a set of video/audio recordings, a textual aligned bilingual corpus, etc.) and (b) metadata about constituents of such Language Resources, e.g. lexicon entries that could be lemmas or words, a single audio/speech file, a single video frame, a sentence of an aligned textual corpus, etc.

For instance, for the LR contents but also documentations and metadata, most of the users are able to produce data (characters) that are encoded in Unicode (or the associated ISO standard 10646). Some of the languages still have to develop their own Unicode official versions. Another example is the Country codes that have been standardised as ISO 3166 and which are *"intended for use in any application requiring the expression of current country names in coded form"*. As with many ISO standards, some controversial debates have arisen about the notion of "official". In addition to the country code, ISO also adopted a "first" list of language code elements (ISO 639) *"comprising three-letter language identifiers for the representation of names of living and extinct language families and groups"*. See the language section about this issue in this report.

The coding aspect is often related to the description of data entries and/or data sets. For instance, some focus on particular types of resources such as the TEI (Text Encoding Initiative) for text encoding, SynAF for Syntactic annotation framework, LMF the Lexical markup framework, etc. More information is given in the coming sections.

Similar standards exist for audio and video resources, in particular the well known MPEGs (The Moving Picture Experts Group) which was formed by the ISO for *"the development of standards for coded representation of digital audio and video"*. MPEG is to be seen as a set of or a family of standards (from MPEG-1 for CD or MP3, MPEG-2 for Digital Television or DVDs; MPEG-4 for multimedia for the fixed and mobile web; MPEG-7, the standard for description and search of audio and visual content; etc.). Other best practices are still in use for instance in the speech domain, where most of the Europeans use SAM format (defined within the European project or NIST/Sphere practice[1]).

In addition, the International Telecom Union works its own standards in particular to code the waveforms transmitted along our telephone fixed lines (these use PCM A-Law in Europe and in other continents while the PCM-µ-law is used in the USA and Japan) but also the GSM, UMTS digital coding. Many applications exploit phonetic transcriptions (phonetic lexica) such as Speech recognition, speech synthesis, voice conversion, etc. Both IPA (International Phonetic Alphabet) and SAMPA and X-SAMPA (SAM Phonetic Alphabet and eXtended SAMPA) are used. The SAMPA (and X-SAMPA) are maintained at the UCL phonetic lab since the end of the SAM project[2]. The adoption of a best practice like this one and reporting on it both at the Metadata level and at the data level help boost interoperability and sharing of the data.

Using some of these standards one can describe a lexicon that comprises 20,000 entries (lemmas, stems, etc.) with additional metadata features such as a descriptive item to report on the context in which the data was produced, to describe the applications for which it was produced, the applications for which it has been/is used, etc. The same can be done for speech data resource, which gather speech for a given language from hundreds of speakers channeled via GSM or fixed lines, for Video data sets and/or combination of all these resources.

The other granularity is about the constituents or data elements which are "an atomic unit of data" (a word/lemma in a lexicon, a speech sample in an audio file, etc.). Several initiatives are on-going to represent the atomic units (e.g. Data Category Registry or Iso Categories, ISOCat).

METADATA is a descriptive system that should reflect the use of such standards / best practices by the data producers and/or providers.

A good number of initiatives, described in the previous sections, worked at defining the key dimensions that help identify and describe Language Resources. From those initiatives,

[1] SAM and NIST/SPHERE differ in their storing information: within the speech file (NIST/SPHERE) or outside the speech file (SAM). The NIST/SPHERE format is supported by NIST and DARPA while SAM (a European `standard'), is used by many players in Europe (including those who exploit Speechdat resources).

[2] http://www.phon.ucl.ac.uk/home/sampa/

different metadata sets were produced. Let us recall some of the important existing metadata sets with the institutions behind them:

- OLAC (*Open Language Archives Community*)[1]: a community for the creation of a virtual library of international Language Resources, with very rich discussions on metadata development and improvement.
- ELRA Catalogue of Language Resources[2]: a catalogue which gathers over 900 Language Resources described in a formalized way, implemented by the European Language Resources Association (ELRA),
- IMDI (*International Standards for Language Engineering Metadata Initiative*)[3]: an initiative for the standardization of metadata used to describe Language Resources,
- LDC Catalogue[4]: a catalogue of Language Resources mainly produced by the LDC (Linguistic Data Consortium), some of which coming from projects funded by the USA government.
- SHACHI[5]: SHACHI is the database developed by the NICT (National Institute and Communications Technology) and Nagoya University aiming at collecting data information on language resources in Asia and Western countries.

Within those different metadata sets, we can highlight a number of key elements identified as important for the classification of LRs. Most of the time those elements are common to the different metadata sets. Some of them are given below.

- Languages:

Information about the language that the LRs cover is by definition essential for the use of Language Resources. Many lists of languages are available. However, to encode languages, the widest agreed upon convention is the ISO standard[6] (ISO 639 for languages and ISO 3166 for countries where languages are spoken).

- Domains & Topics:

In many technologies that are data-driven (in particular the statistical approach to systems training), it is essential to acquire data that is representative of the specific domain being tackled and even better to focus on some specific topics (or sub-domain) within such domain.

A number of surveys attempted to identify the major domains represented in the digital world (the digital content) and the prevailing ones are Tourism, Commerce/Trade, Health-Medicine, Education, Law, IT/telecommunications, Administration, etc. Most of these domains have now the corresponding e-activities (e-Tourism, e-Health, e-government, e-Learning, e-Commerce, etc.).

---

1 http://www.language-archives.org

2 http://catalog.elra.info

3 http://www.mpi.nl/IMDI

4 http://www.ldc.upenn.edu/Catalog

5 http://shachi.org/shachi/en/index.php, (*Shachi* means “orca” in English).

6 http://www.iso.org/iso/home.htm

Many of current research projects focus on these prevailing domains as long as they are lucrative through their support to multilingual and multicultural business and also reflected in the availability of the "data" through unrelated production processes, (data being texts, audio recordings, video, etc. produced for the exploitation of such domains and re-used to develop/enhance the corresponding technologies). Let us illustrate some of these domains:

- Tourism:

Tourism is a domain that, according to the findings of previous surveys, is significantly represented in multilingual text production (availability of electronic handbooks, phrasebooks, audio recordings, terminologies and dictionaries, etc.). This potential interest is confirmed by several related projects funded by the EC and/or national programmes (in several programmes e.g. e-content, FP etc. such as eMate, FAME, M-GUIDE, LC-STAR, MEDIA, NESPOLE, etc.). The multilingual aspect of this domain promotes the use of multilingual technologies such as Machine Translation, speech-to-speech translation, multilingual dialogue, etc.

- Health:

Health-related information is among the most frequently accessed types of information on the Internet. The increasingly used term "eHealth" generally denotes the combined use of electronic communication and information technology in the health sector and their use within the professionals and/or for the general public. A large number of projects, in particular related to Information extraction, retrieval, and question-answering, focus on this domain.

- Education:

The domain of education is one of the major interests of the EU. It should be noted that there is a strong tendency to adapt the EU's education and training systems to the knowledge economy and digital culture. Unfortunately, few applications encapsulating language technologies are exploited (e.g. language learning). The eLearning initiatives seek to mobilize the educational and cultural communities, in order to speed up changes in the education and training systems for Europe's move to a knowledge-based society. Some key research into innovative eLearning solutions is enabling learners to use existing, widely available technologies.

- Law and related information:

Most of the Machine Translation research, worldwide, focuses these days on the legal information that is provided by the European institutions (e.g. EU Parliament). In addition, Legal information constitutes a significant part of public sector information at the national levels. The multilingual aspect is dealt with by the multilingual countries (Canada, HongKong, Spain) and at some degree by the international agencies such as those of the United Nations (e.g. general assembly, UNESCO, UNICEF, etc.) or some multinational companies.

- Domains and Language Resources:

All the domains described above are more or less represented in the existing LRs but the percentages are still low compared to the LRs describing the general language. We should also take into account that the majority of the LRs, that describe either the general language or a domain specific, are monolingual. This is attributed to the fact that a significant number of the resources are nationally funded, and national funds favor general language over domain specific resources. Previous finding have also attested that domain specific resources are mainly funded by end users (industrial funding). A number of national initiatives conducted serious plans to produce a national corpus (some even a national basic LR kit), hence we can find today the *British National Corpus*, the *American National Corpus*, the *Polish*, *Croatian*, *Czech National Corpus*, *Hellenic National Corpus*, *Hungarian National corpus*, etc.

Regarding the Topics, we can easily understand that referring to tourism or law (in our previous sections) is too general and may still require particular customization (tourism may need to focus on hotel bookings while law refers often to available parallel corpus containing European Union (EU) documents of mostly legal nature.

- Time span:

In many sectors of activities, it is important to have access to data that reflects the use of a language during a period of time. For instance, all cultural heritages require the use of specialized dictionaries/terminologies/ontologies/etc. that may consist of the use of terms that are obsolete nowadays. This is often the case for the "old" Press. It is crucial to collect such texts to train both Information retrieval systems and Machine translation ones to handle such cultural and/or data collections.

- Formats:

We have already touched upon format in our description of Metadata and standards, but we feel that it is important to highlight some other aspects herein.

Data format is the primary obstacle to sharing data and rendering them interoperable. The format often refers to storage and encoding techniques: somehow to the "container" (files, databases, etc.) and to the "content" (the list of entries, sequence of words, individual speech or video samples, or any annotation related to such items, etc. that also addresses the encoding of their basic constituents such as the encoding of characters (e.g. Unicode[1]), of speech samples (16 bits), etc).

The most important work with respect to that was carried out in the framework of a number of European projects such as EAGLES[2] (*Expert Advisory Group on Language Engineering Standards, 1993*) that defined standards for corpora regarding their typology, encoding, and annotations (syntactic and morpho-syntactic) leading to the CES[3] (*Corpus Encoding Standard*) as an instantiation and follow-up of the TEI[4] (Text Encoding Initiative). A

[1] Unicode (also ISO-10646) is now THE standard to character encoding, though some issues are still pending for many languages that have poor ICT (&HLT) communities.

[2] http://www.ilc.cnr.it/EAGLES/home.html

[3] http://www.cs.vassar.edu/CES

[4] http://www.tei-c.org

working group within EAGLES also elaborated on speech resources (the famous "Spoken Language Working Group", WG5), and came up with recommendations referred to as "Working Standards" to allow for the production of large speech corpora, also based on the effort that were devoted earlier to speech resources in the framework of the SAM ESPRIT project (1987) that came up with the SAM format, widely used in Europe while at the same time, US researchers adopted Sphere format within the DARPA funded projects. SAM was followed later on by a family of projects known as Speechdat-Family that elaborate a stable model for speech data production processes including data exchange format but also efficient validation and quality assessment.

EAGLES based a lot of its work on SGML and later on adopting XML as the approach to represent (mostly) semi-structured data and help its processing by widely available tools. Some extensions have been developed to help quickly deploy applications including for speech technologies (VoiceXML).

Many other projects addressed specific format needs so as to reflect aspects such as alignment of corpora, multilingual lexica, etc. For instance, Parole and Simple projects continued the implementation of EAGLES recommendations with respect to syntax and semantics.

The major problem with these formats (and to some extent the standards used) is that they do not consider enough community practices and we still feel that this area of language technologies is just coming to some maturity.

3.2.6.1 Coordination and Harmonization initiatives

Within FlareNet and SIL, there is a strong willingness to work on the harmonization of the LR catalogs of the largest international data centers (in particular the ones mentioned in these sections of ELRA, LDC, as well as the Shachi union catalog produced by NICT). The approach to this harmonization will rely on the production of a super-set of the metadata elements used by each of the partners. More on this initiative will be announced at the next FlareNet Forum planned for mid February 2010. The aim of such harmonization is to allow efficient and accurate searches over all existing LRs knowledge bases.

3.3 LRs and the BLARKS

3.3.1 The BLARK concept and ongoing initiatives

The Basic Language Resource Kit (abbreviated BLARK) is a concept introduced by ELSNET and ELRA to boost the views on the minimal set of language resources that is necessary to do any *pre-competitive* research and education[1]. As a concept it needs a careful monitoring of how it can be applied to a given area overtime and also requires a strong attention to the state of art and its evolution within the R&D and technology activities.

[1] *ELSNET and ELRA: A common past and a common future*, S. Krauwer, ELRA newsletter, Vol.3, N°2, May 1998

A BLARK comprises many different elements, such as:

- Basic language resources, such as written and spoken corpora, dictionaries, grammars
- Basic tools, such as taggers, parsers, annotation tools, speech/image/video features extractions
- Benchmarks (test suites, metrics, methodologies) for evaluation
- Etc.

The underlying idea is to wisely agree on a common generic BLARK definition, based on the collective experience and expertise gained with many different languages by the members of the Human Language Technology community at large. This common definition will save time and effort when defining the needs and requirements of a given language and/or a given community of research. It will allow for porting of knowledge between languages, it will ensure interoperability and interconnectivity (especially for multilingual or cross-lingual application areas), and it will help making realistic estimates of costs and efforts required to produce them.

The definition is in principle intended to be language independent, but as specific languages do come with different requirements, instantiations of the BLARK may vary in some respects from language to language. Once the definition for a specific language is available, it can be decided which of the components are already available and which ones are missing. The amount of missing components may vary dramatically from language to language, as, for instance, some of the major languages such as English may already be fully covered, whereas others may have to start from scratch. Once the gaps are identified, priorities have to be assigned to the components to be produced, in order to make a realistic plan for the gradual completion of the BLARK. At each of these stages, a consensus on the needs but also a compromise on the items to produce versus their costs should be sought.

The BLARK concept was first presented by S. Krauwer (1998) in the ELRA Newsletter. After this publication the idea has been taken up by the Dutch Language Union (DLU), the intergovernmental body created by the Dutch national and Flemish regional government to take care of their common language. A number of publications have followed from these activities, describing both the result (a fairly concrete enumeration of components that should be included in the BLARK for the Dutch language) and the process that led to this result. An excellent summary of the process and the results of the Dutch BLARK exercise can be found in an article by Binnenpoorte et al (2002) in the proceedings of the LREC 2002 workshop "*Towards a Roadmap for Multimodal Language resources and Evaluation*" organized by ELSNET.

To conclude, the BLARK concepts led to inventories (and possibly discussions on existing gaps) that are derived substantially from existing catalogues (e.g. ELRA and LDC) but the problem that remains is how to go beyond the publicly available and/or visible resources.

3.3.2 BLARK matrices

ELDA developed a BLARK[1] interactive web service that allows to consolidate the identification of the needs in terms of LRs. This service enables to cross-link LRs with expected applications and possible languages.

ELDA implemented a matrix with a list of applications and potential modules that could be linked to LRs and languages. The resulting matrices were developed in part thanks to the works carried out within the NEMLAR project (see following section). Two distinct matrices were developed and can be filled in directly through the web site:

The first matrix "Applications/Modules" (see below) allows users to insert a mark giving (to their point of view) the level of importance that needs to be considered per module with respect to a given application. This can be provided both for Speech and Written language areas: important (✚), very important (✚✚), essential (✚✚✚) or not applicable (⌀).

Spoken applications vs **spoken modules** for **Arabic** language

close window	Customization to Different	Dialect/Language	Dictation	Embedded Speech	Emotion Identification	Emotion/Prosody Output	Generation Lips Movement	Lips Movement Reading
Acoustic Models	✚✚✚	✚✚✚	✚✚✚	✚✚✚	✚✚✚	✚✚✚	✚✚✚	✚✚✚
Dialect/language Identification		✚	✚	✚	✚			✚
Emotion Identification		✚	✚	✚		✚✚		✚
Language Models		✚✚	✚✚✚	✚✚		✚✚		
Lexicon Adaptation			✚	✚				
Lips Movement Reading		✚✚						✚✚✚
Phoneme Alignment			✚	✚				
Pronunciation Lexicon			✚✚✚	✚✚✚				
Prosody Prediction						✚✚✚		
Prosody Recognition		✚	✚	✚	✚✚✚			
Segmenter Speech/silence		✚✚	✚✚	✚✚	✚✚	✚		✚
Sentence Boundary Detection		✚	✚	✚	✚✚	✚✚		✚
Speaker Adaptation		✚	✚✚	✚✚	✚			✚
Speaker Recognition/identification		✚	✚	✚	✚			✚
Speech Units Selection						✚✚✚		
Speech/non-speech Music Detection		✚	✚	✚	✚✚			✚
Word Boundary Identification		✚	✚	✚	✚	✚✚		✚

Figure 1: Abstract from the "Applications/Modules" matrix for Arabic language, in the Speech area:

The second matrix, "Resources/Modules", uses the same marking method (important (✚), very important (✚✚), essential (✚✚✚) or not applicable (⌀), but to link LRs needed to specific modules.

[1] http://www.blark.org/

Spoken resources vs **spoken modules** for **Arabic** language

close window	Annotated Written Corpus	Audio Data with Prosodic Markers and other	BNSC	Desktop/Microphone & High Quality	Non Vowelised Corpus	Onomastica (proper names)	Phonetic Lexicon	Telephony
Acoustic Models		✣✣✣	✣✣✣	✣✣✣				✣✣✣
Dialect/language Identification		✣	✣✣	✣✣		✣	✣	✣✣
Emotion Identification		✣	✣	✣		✣	✣	✣
Language Models	✣✣				✣✣			
Lexicon Adaptation	✣				✣	✣✣✣	✣✣✣	
Lips Movement Reading								
Phoneme Alignment	✣✣	✣✣	✣✣	✣✣		✣✣✣	✣✣✣	✣✣
Pronunciation Lexicon	✣					✣✣✣	✣✣✣	
Prosody Prediction	✣✣	✣✣				✣✣	✣✣	
Prosody Recognition	✣✣	✣✣✣		✣		✣✣	✣✣	✣
Segmenter Speech/silence		✣✣	✣✣	✣✣				✣✣
Sentence Boundary Detection		✣✣	✣✣	✣✣		✣	✣	✣✣
Speaker Adaptation		✣	✣✣	✣✣				✣✣
Speaker Recognition/identification		✣	✣	✣				✣
Speech Units Selection	✣✣	✣✣✣		✣		✣	✣	✣
Speech/non-speech Music Detection		✣✣	✣✣	✣				✣
Word Boundary Identification		✣	✣	✣		✣	✣	✣

Figure 2: Abstract from the "Resources/Modules" matrix for Arabic language, in the Speech area

In the following sections, we will focus on a number of languages and particular technologies to better emphasize the importance of BLARK for all the actors involved.

Although there exist ongoing initiatives around the BLARK concept for European languages, those are still few. Worth-to-be-noted initiatives and projects are the following:

- NEMLAR ((http://www.medar.info/The_Nemlar_Project/, 2003-2005), and later on MEDAR(http://www.medar.info/index.php, 2008-2010), two projects supported by the European Commission's ICT programme that addresses International Cooperation between the EU and the Mediterranean region on Speech and Language Technologies for Arabic. The NEMLAR project has elaborated the first BLARK for Arabic, which, taken together with the results of the survey conducted on existing LRs, was a very good starting point for deciding on priorities for development of specific LRs. MEDAR has a specific task on reviewing the Arabic BLARK, established within NEMLAR and updating it considering the emerging needs and the newly identified resources and tools. MEDAR has carried such updates on BLARK with a focus on MT and MLIR.

- STEVIN programme[1] (Dutch acronym for "Substantial language and speech technology resources for the Dutch language"): The programme is lasting 7 years (from 2005 to 2011). The aim of this programme is to stimulate the language and speech technological sectors in Flanders and the Netherlands, in order to enhance the innovation capacity of this sector as well as the position of Dutch in the modern information and communication world. In relation with the Dutch language, it aims at promoting research in Human Language Technology (HLT) and developing Language Resources, raising awareness and demand of HLT product, and organizing the management, maintenance and distribution of developed resources.

- "An Infrastructure for Swedish language technology" Project: This project was financed by the Swedish Research Council in 2007-2008. Part of the project focused on a BLARK for Swedish[2]. For instance, in May 2008, SALDO, a full-size Swedish lexical resource was made available under an open-source license by Språkbanken, University of Gothenburg. Containing more than 50,000 lemma entries and being freely available, SALDO fits the requirements of the morphological lexical component of a Swedish BLARK.

- Towards an Icelandic BLARK: some projects are emerging to fill the gaps for the Icelandic language. One of them is known as IceNLP[3] and aims at developing a Natural Language Processing toolbox, IceNLP, for analyzing the Icelandic language.

In addition to these known BLaRK per language, mostly European-based and supported, let us mention the Less Commonly Taught Languages (LCTL) project conducted at LDC, which is an excellent exploitation of the BLaRK concept from our point of view.

"On the contrary of other descriptive initiatives, the goal of this LDC project was to create and share resources to support additional basic research and initial technology development in what have been called Less Commonly Taught Languages. These languages have also been called Low Density, not for the population of native speakers but rather for the scarcity of resources. A typology that distinguishes both population of native speakers might label them High Density/Sparse Resource language since the languages of current focus have more than a million speakers but inadequate resources for building human language technologies".

LDC planned to collect resources that would allow developing some basic technology up to SMT. To achieve such endeavor, the resources list had to cover:

- Monolingual Text: preferably news text written originally in the LCTL. Monolingual text will be tagged, tokenized and converted into a standard encoding, about 250 Kwords that will be translated into English plus an additional 250 Kwords.
- Parallel Text: preferably news text written originally in the LCTL aligned with English (including exploiting monolingual texts and have each source sentence translated into one or more sentences in the target language), around 175 Kwords

[1] http://hmi.ewi.utwente.nl/project/STEVIN

[2] http://stp.lingfil.uu.se/~bea/blark/home-en.html

[3] http://nlp.ru.is/projects.htm

translated from the LCTL into English plus 75 kwords translated from English to the LCTL.

- Bilingual Lexicon: containing a minimum of 10k lemmas but targeting larger lexicons that provide 90-95% coverage over the monolingual text corpus.
- In addition, LDC ensured that a number of basic tools are developed and/or adapted to these languages, e.g. Encoding Converters (converting raw text and lexicon encodings to Unicode); Word and Sentence Segmenter (having in mind that many languages do not already show word segmentation in their writing systems and/or do not mark sentence boundaries explicitly); POS Tagset, Morphological Analyzer, Named-Entity Tagged text and the tagger, Personal Name Transliterator, etc.

Among the languages addressed in this project, we can quote Tamazight, Urdu, Thai, Hungarian, Bengali, Punjabi, Tamil, Yoruba, etc.

All over the world and essentially in Europe, similar initiatives are being conducted at individual labs without too much concertation and definitely with no coordination at national/regional levels.

In addition to the pure definition(s) given above, one should also consider a number of attributes such as those introduced by ELRA within its QQC (see above). Many of these attributes are related to availability of the resources composing the BLARK, the "quality" of such components, the size of each component and the expected performance one can achieve by such quantity of data, the interoperability issues with other components (and possibly the use of standards, etc.).

Many of these notions are very hard to define in absolute terms. Availability (in term of costs) of a German BLARK for 1000€ may be easily understood by the European research community while a similar cost for a Swahili kit (assuming it exists) will certainly hinder the use by the local research community even the one based in Europe that may not be willing to invest on a PhD work at that price.

Quality is another (difficult) concept to assess. It can be relative to the data documentations (accuracy for instance) but it is more critical when it addresses issues relative to the content. In the NEMLAR project, we defined at least four criteria of paramount importance; these were: the extent that the resource is based on a common standard, the way the resource is based on well-defined specifications, the way the resource is suitable for a specific task and finally how the resource is interoperable with other resources and tools. Of course, these attributes are not completely independent and one can easily add a few more.

As illustrated by the LDC project, it is crucial to provide quantitative figures for the various resources needed: how many words in a corpus, how many hours of speech, how many parallel texts and lexical entries for SMT, etc. It is important that a second stage of the BLARK definition includes guidelines for what counts as a sufficiently large corpus, lexicon, etc. Two approaches are debated herein.

The first one considers general costs and all related problems to come up with a reasonable size of a corpus, lexicon, etc. Such approach was adopted as seen above within the LDC LCTL project. The second approach, advocated for by ELRA and many of its partners, is to

search for the best compromise between quantity of data used and the performance to be achieved by a baseline of the given technology. Of course, at the end, the same problems of specifying and funding the production arose (more details on specific technologies in the corresponding sections).

3.4 Examples of BLARKs for some key applications: requirements, costs and technology performances

In the following paragraphs, we will briefly introduce some key technology areas with their associated BLARK matrices. The idea is to highlight the diversity of required resources, and some crucial cases emphasize such requirements through some figures on the size of data expected but also, when available, stress the performance one should expect from a baseline or a well known system given the data sizes.

3.4.1 Example of Cross-lingual Information Extraction, Retrieval & Question-Answering

The following 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[1]. We start our BLARK section with this technology to illustrate the matrices we introduced above.

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 on the left hand side of the two tables.

[1] For details please refer to the work done in Treble-CLEF project (http://www.trebleclef.eu/publications.php) and in particular to the deliverable D5.2 - Best Practices in language Resources for MLIA and D3.1 - System Developers Workshop.

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: Resources vs. Modules

Here is a list of the most needed language resources for developing and running a CLIR system (with some indications of typical examples and data size):

- Stop word list: Stop word removal consists in removing words that are useless for the information search task. This is done using pre-defined and language-specific stop-word lists (e.g. the English stopword list used for Snowball stemmer comprises around 250 words). Most languages have such list freely available.

- Corpora: corpora are necessary to the development of different NLP components (morphological analyser, Pos-tagger, etc.) that are essential to IR system. When IR systems are developed in a cross-lingual context, parallel multilingual corpora are essential. They are used for the development of statistical MT modules (query translation, MT of retrieved documents or answers). Some examples of efforts required are given in the section dedicated to the MT technologies for the un-annotated data. The morphological annotation is not yet a solved problem for many languages and still require human efforts to annotate huge corpora for training (or to correct semi-automatically annotated corpora). Efforts depend on the languages.

- Lexicons: A lexicon gives the vocabulary of a specific language, including its words and expressions. In the CLIR context, multilingual lexicons constitute a key resource; it is generally the corner stone of a query translation module (e.g. a bilingual English-French dictionary as EURADIC contains around 250,000 pairs of French-English terms, with their part of speech for 8K€).

- Grammars: Grammars are models of the linguistic structure of languages. A grammar is necessary for syntactic parsing which is the process of analyzing a sequence of tokens to determine its grammatical structure and major functional relations between words. Grammars are embedded in the syntactic parsers for

which they were developed (see for instance the free for download Stanford parsers[1]).

- Thesauri and ontologies: 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). In the CLIR context, multilingual ontologies are needed. The nature and size of ontologies closely depends on the domain of the IR application (science, medical, politics, etc.). As an example, the well-known multilingual thesaurus for medical terms, MeSH[2], contains around 160,000 entry terms and uses 25,000 descriptors (version of 2009) and is freely available in many languages (equivalent CISMEF for French).

3.4.2 Example of Statistical MT, required LR and related costs, performances

The work carried out within the EC funded project MEDAR elaborated a detailed plan for the production of a prototype of Statistically-based Machine Translation Engine. More has been done with the TC-STAR project (http://www.elda.org/article.php3?id_article=166) and also, more recently by the EuroMatrix (and EuroMatrix+) projects.

The needs were expressed along the lines of the BLARK defined features and what we need is exemplified herein with some comments regarding the cost of production/availability.

- Bilingual corpora, size 20–50 Million words; no one expects these to be derived from commissioned translations for this purpose (the cost would be 3M€ for 20Million words (0.15€/word x 20Million) unless it is made with serious public funds. In most cases this is achieved thanks to existing translations from internet/archives within organizations that use such languages (example: United Nations for its official languages Arabic, English, French, Russian, Chinese; Other United Nation agencies such as UNESCO, FAO, etc; that often produce bilingual documents; the European Union for its official languages that led to the EuroParl / JRC corpus, etc., Press agencies that often produce multilingual press-release of small size, etc.). In this context it is important to assess the efforts required to identify the sources, to negotiate the rights to use the data for this purpose, to download/crawl the data, to clean it, to align it, to verify and validate the quality of alignment which is an essential parameter for the quality of translations, and to run the SMT package to learn our translation statistics.

- Monolingual corpora of the target language: this is essential for the quality of the generated texts and usually developers expect about 100M words; Obviously one can get more easily such monolingual corpora but this depends on the target language presence on the Internet. Many languages are still facing the digital divide and have

[1] Stanford syntactic parsers: http://nlp.stanford.edu/software/lex-parser.shtml

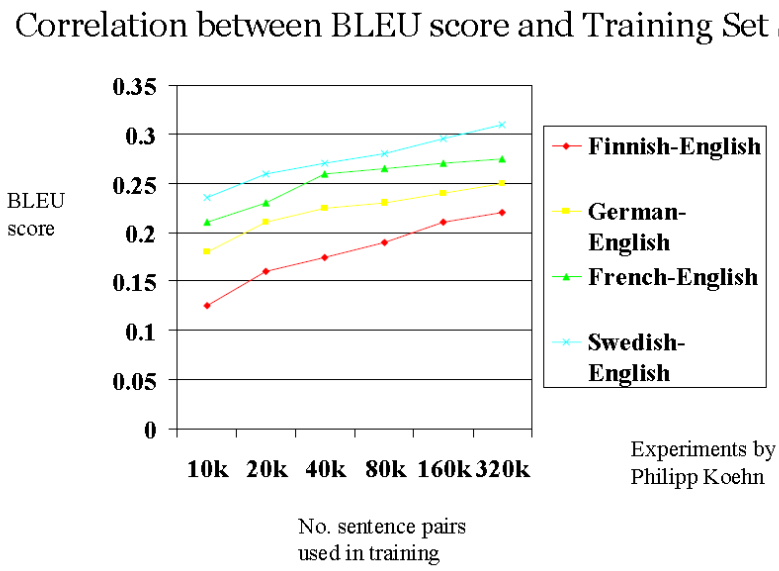
[2] MeSH (National Library of Medicine's thesaurus): http://www.nlm.nih.gov/mesh/

no “texts” on the internet (for instance some languages that have been normalized in terms of writing system just recently or those used by minorities).

- We can imagine a small “high quality” corpus for evaluation (1 reference & 4 translations carefully checked) of 20K words.

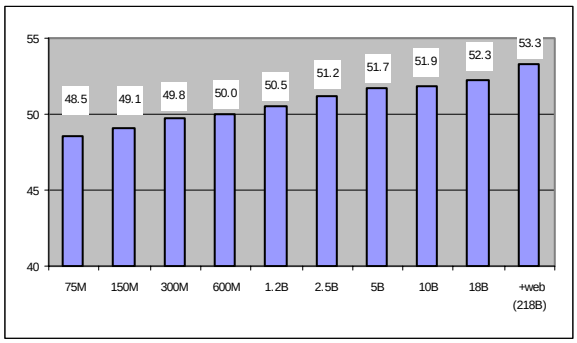
Of course, SMT is not the sole approach to Machine Translation. Other more traditional approaches (e.g. the conventional rule-based methods) are still in use and new hybrid approaches are experimented. For instance, in order to develop a grammar-based and transfer approach for Catalan, one needs to develop a specific MT lexicon that requires about 2p*y of a highly skilled computational linguist.

As stated above, the data sizes given above are estimate figures to help set up terms and limits. If we want to adopt the performance versus size of the data, we will end up with rather different figures. Let us illustrate this through the following diagram, courtesy of Andy Way/Hany Hassan (DCU):



According to Philip Koehn, “*doubling the training data gives constant improvement of +1% according to the BLEU measure”*. How the Bleu measures are correlated with human judges is somehow agreed upon but its correlation with the “usability” aspects is still debated.

Another illustration of this aspect from Franz Och (Google) is given herein with the improvement obtained while doubling the training data for Language Modeling (the +web 218 Billion words are out-of-the vocabulary data, unknown language pair, very likely English-Arabic, presentation at TC-STAR public workshop, 2005).



3.4.3 Example of Automatic Speech (broadcast news) transcriptions

The same concept can help us derive the needs and requirements of other technologies in the speech arena.

For instance to produce a transcription engine for audio data from broadcasted news (to convert speech to text), one needs to produce:

- 100h of manually transcribed broadcast recordings (with as much variety as possible)
- a phonetic lexicon covering all words and expressions (typically 64k entries or more)
- 100 million words (or more) of textual data for language modeling purposes

In addition, one needs software/algorithms to implement:

- A speech parameterization front end
- A HMM training toolkit like HTK or Sphinx
- A language modeling toolkit like SRILM or CMULM
- A multi-pass speech decoder like Julius or Sphinx decoder

If we consider the work carried out by the Spoken Language Processing Group at LIMSI/CNRS (http://www.limsi.fr/Scientifique/tlp/), we can illustrate the performance, measured as WER (Word error rate), over the number of hours of audio data used for training as follows (Courtesy of Lori Lamel [1]).

[1] Lori Lamel, Jean-Luc Gauvain, and Gilles Adda. Unsupervised Acoustic Model Training. In Proceedings of ICASSP, pages 877-880, Orlando, May 2002.

Two different techniques are tested (light supervised vs Supervised), the language is US English. More details are given in that paper but also on a paper by (Thomas Pellegrini & Lori Lamel)[1] focusing on a less resourced language, Amharic:

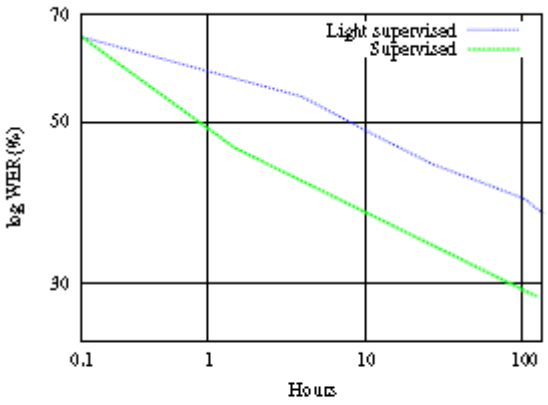


Fig. 1. Word Error Rate (%) as a function of acoustic training data quantity (taken from WERs reported in [2])

3.4.4 Other speech processing technologies

For a **Speaker verification system**, the Language Resources required are a 500 speakers database with recordings of words and sentences, each database manually annotated with the speaker identification label (from 30s of speech to a few minutes) and the Software/Algorithm that implements:

- A speech parameterization front end
- A HMM/GMM/ANN model training toolkit
- A speech verification software implementing a DTW, VQ, HMM/GMM or ANN approach

For **Facial recognition**, the Language Resources required are a large set of digitized images of human faces, annotated at various levels (association of identifier with each image, face bounding boxes, eyes, nose, etc). One also needs software/algorithms that implement:

- A feature extraction front end
- A model algorithm using HMM/ANN/eigenfaces technique or others
- A recognition software algorithm using HMM/ANN/eigenfaces/other approaches

3.4.5 Conclusions about BLARKs and its various interpretations

[1] Thomas Pellegrini and Lori Lamel. Are Audio or textual training data more important for ASR in less-represented languages?. In 1st International Workshop on Spoken Languages Technologies for Under-resourced languages (SLTU), pages 2-6, Hanoi, Vietnam, May 5 - 7 2008.

The BLARK is a useful instrument, supported by many players (including researchers & funding agencies), to assess the coverage of a given language by current language technologies. It is also useful for roadmapping and planning, assuming data is collected accurately.

In order to do so, it is highly recommended to describe the HLT scene for each given language, which requires a harmonized description tools as described herein. It is also important to avoid misleading conclusions regarding availability of the BLARK for a given language; In addition to its costs, one should be very careful about the performance one may achieve with such resources. In addition to that, policy makers should use the instrument wisely to detect and monitor the emerging needs and the corresponding novel research themes and innovative applications. This, of course, requires strong assessment of the state of the art, based on evaluation paradigms as implemented within community supported evaluation campaigns.

Such approach could help establish current technological "baselines" per language, per technology, etc. with a clear picture of important barriers and threats.

It is also important to monitor the evolution of existing resources along various dimensions over time; we may list as key factors availability for a given language, its technical characteristics, etc. For instance if we consider the Broadcast news corpus, one can monitor availability of such resources for all EU languages, the size (in terms of number of hours of audio data), the annotations produced (orthographic transcriptions, part of speech tagging, etc.), etc. Such monitoring will be conducted through the LREC Map as of LREC'2010.

The various dimensions mentioned herein (Language coverage, modalities, domains, etc.) require efforts that can be only guaranteed by strong involvement of international, national, regional, transnational agencies in a shared effort and may be involving more Joint Public/Private partnerships. A dedicated **Investment fund** should be carefully envisaged (under current economic situation), at least for the coordination and promotion of Best Practices, Standards, Interoperability.

3.5 LRs sharing conditions and principles

Even if the main purpose of HLT organizations (should it be public or private) is not to distribute Language Resources *per se*, we have reached a point where the HLT field needs a sharing of such resources, for various reasons:

- Participation in the market evolution
- Exchange of information
- Return on investment
- Capitalization on previous work

The activity of LR distribution (and sharing) is not easy and requires the implementation of a number of services and policies. Some of the most critical ones can be listed as:

- Clear view of the technical aspects
- Typology of use, typology of users

- Pricing policy: Fees versus free
- Legal issues: license types but also ethics, privacy, etc.
- Distribution issues: media, confidentiality, reliability, integrity, etc.
- Usability, accessibility, interoperability and scalability, and overall sustainability.

3.5.1 Technical and logistic requirements:

- Archiving the LRs :

This first step is of great importance to prevent the loss of data along the years. Only a systematic archival can prevent from a definitive loss. Beyond the risk of loss, it is of utmost importance to maintain an organized archive enabling to easily find and reproduce the LRs whenever needed. A following up of technological evolution is also required to make possible the change of media for distribution along the years.

- Detailed information about LRs:

As described in section 3.1, it is required to provide to any potential user detailed information on the distributed LRs. The help of metadata sets and the availability of detailed catalogues are therefore needed.

- Duplication of LRs:

Appropriate technical material is required to prepare the LRs for distribution. Current electronic media in use are mainly CD-ROMs, DVD-Roms or hard disks, which requires using a number of different materials, such as duplication material, labeling, packaging, etc.

- Delivery:

At the final stage of the distribution cycle, delivery means and costs need to be estimated, as well as delivery time, insurance, packaging, etc.

3.5.2 Legal issues:

In order to allow a provider to distribute its resources to any user, legal issues must be taken into account. With Legal issues, we very often confuse Legal ownership (IPR, Copyright, author's right, the EC Database directive, etc.), Ethical, Privacy and other related issues. Developing such issues in detail is not part of this deliverable[1].

- Obstacles to be pointed out here are various:

The lack of legal concern within academic centers, who "omit/forget" to ask for prior authorizations, and clean IPR, prior approval, etc.

- The use of legal models with unduly restrictive distribution rights:

The different strata of intellectual property rights, which are not taken into account (e.g. the production of new resources that integrate resources already covered by intellectual property rights);

[1] A Guide for the Production of Reusable Language Resources, Victoria Arranz, Franck Gandcher, Valérie Mapelli and Khalid Choukri, LREC 2008 Proceedings

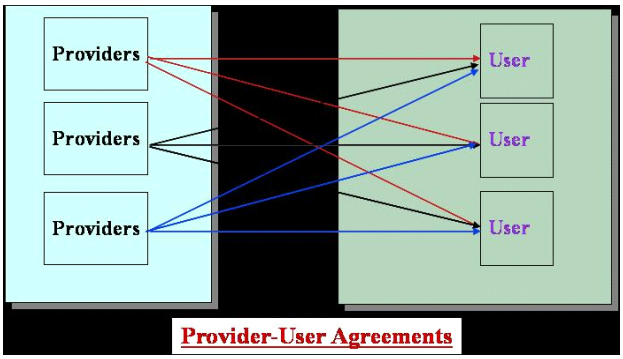
- The multiple home-made license models, often inspired by software license models (such as GNU, GPL, Creative Commons), which are generally not adapted to Language Resources.
- The diversity of legal protection modes in Europe and over the world.

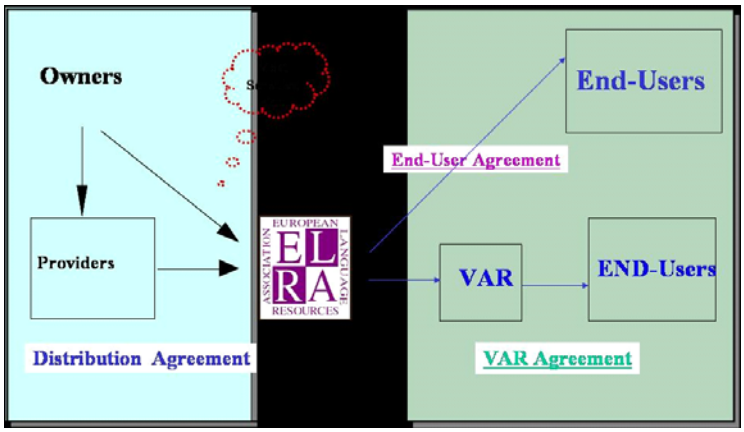
When considering a LR, we have to distinguish different types of players: (1) rights owner(s) –of one portion of the resource (e.g. a speaker who gave his/her voice in the case of speech resources) or of the whole resource, (2) providers (which may be the owner itself), (3) distributors, (4) application integrators/developers and (5) end-users (R&D teams). With the aim of making a resource available, one needs to define the relationship between those different players, through adequate legal agreements, at the proper time.

Within a project financed by the French Ministry of Research, ELDA wrote a guide aiming at offering technical, legal and strategic recommendations/guidelines for the reuse of Language Resources. As far as legal guidelines are concerned, ELDA has drawn up an inventory of different licenses, which are being used in the field. The fact that most existing licenses are « home-made », and most of the time inspired from software distribution models (such as GNU, GPL, Creative Commons) should be emphasized. Based on that information, and together with ELDA's own licenses developed specifically for the exchange of LRs, ELDA extrapolated a table comparing the rights described in those different existing licenses. Some of the elements of comparison that could be identified are listed below:

Domain of application: one can observe that some licenses can be applied directly to Language Resources, whereas others are inspired from library products, software distribution industry or more fuzzily to any type of works of the mind (IP).
Types of use: limitations of use of LRs are leveled within the different licenses depending on the groups of users that are targeted. Some licenses focus only on education purposes, whereas others may allow research use only or may cover the possibility of integration in distributable/commercialized products.
Modification issues: results obtained from the LRs can be modified following constraints laid down in the license. Some licenses may allow the full modification and reuse of the LRs whereas other may not allow modifying them at all or under certain circumstances. For example, one LR may be used only within a research team for its specific research or even a specific project or activity (e.g. Evaluation). Types of use and modification issues are usually very cross-linked.
Re-distribution: re-distribution right is one of the core concepts of free licenses. For example, the Creative Commons model includes a "sharing" mention, which implies a free re-distribution of the source resources. Other licenses may not give as much flexibility.

The importance of the mediating data centers (middleman) in this area has been highlighted by ELRA through the following two diagrams. The first one indicates the usual one-to-one (or peer-to-peer or bilateral) agreement, which requires a good knowledge of the different actors, different licenses, etc. (though some licenses have been standardized today). The second one assumes the existence of a middleman capable of negotiating distribution rights with the owners/providers and passing them on to the users with acceptable and simple licenses.





These legal issues do not affect neither the sharing and/or distribution of Language Resources nor many processes around. Legal aspects are important to understand e.g. the ownership of the Language Resource's Specification (in particular when these are derived from a customization of an existing product), the maintenance "rights" (corrections of bugs, improvements, updates, etc.), the status when merging (fusion) and integrating in new resources, etc. All these aspects may be dealt with through adequate license or left open to never-ending discussions.

Similar problems concern the Evaluation outcomes (campaigns, packages: data, metrics, dissemination of comparable results, etc.).

An illustration summary is given by the Creative Commons (CC) licenses that base most of their legal advice on national common laws (requiring experts to draft licenses per country, under the assumption that users would know and comply with the provider country common law).

This summary is illustrated by the following picture:

Attribution (by)		Reference to the owner
Attribution Share Alike (by-sa)		CC similar for derived resources
Attribution No Derivatives (by-nd)		No modifications
Attribution Non-commercial (by-nc)		No commercial use
Attribution Non-commercial Share Alike (by-nc-sa)		Example of Combination 1
Attribution Non-commercial No Derivatives (by-nc-nd)		Example of Combination 2

N.B. Share Alike: *If you alter, transform, or build upon this work, you may distribute the resulting work only under the same or similar license to this one.*

Licenses versus freely available resources:

Under the European laws and regulations, there is no exception to the Copyright Law for research purposes as the one establish by the US law (section 107 of the copyright law). ELRA is strongly advocating for the establishment, at the EU level of a similar Fair Use for Research purposes of copyrighted material. A Petition has been initiated by ELRA at the Vienna Forum and will be debated at the next one (planned for Mid February 2010).

3.5.3 Pricing policy:

In many contexts, sharing and/or distributing LRs are seen as a new business opportunity for all kinds of HLT players (providers from R&D labs, specialized commercial organizations, technology developers, etc.). With the establishment of ELRA in 1995, such market (so called Business-to-Business market) has been boosted with more providers being involved. The idea was to promote a fair exchange of LRs between different players, along different policy dimensions: (a) to buy LRs at fair market conditions; (b) to acquire LRs for R&D activities at low (or no) fee; (c) to trade and swap some resources against others with similar features and/or quality; (d) etc. Mostly these scenarios developed nicely.

Nevertheless, it is still very hard for a LR producer to set up the appropriate price of its LRs. Indeed, the complexity of the LR contents and the assessment of its quality are often very subjective and make it difficult to raise a fix price. Very often, the price is based on the production cost or even only on a good relationship with the customer. Different opinion of prices can also be found whether the producer/provider comes from the private or public area: the first will expect a financial benefit whereas the second will look more for the technical benefit.

Prices were set up according to a few parameters and expectations:

(a) on the basis of production costs that the producer would like to recoup (Return On Investment)
(b) on the basis of "fair" market price whenever such price exists and widely agreed upon (existence of similar resources for "comparable" languages, market sizes, etc.)
(c) special offers whenever producers were funded (even partially) by public agencies that managed to impose their pricing policy at the start.

From ELRA's experience in the field of Language Resource distribution, it appears that public organizations, who have already been financed to produce LRs expect to exchange resources for free (to their opinion, the resource was already paid thanks to the funds), whereas private organizations who invested their own time and money expect some return on investment. However, it remains very crucial to bear in mind the real value of LRs, as well as the value of the distribution services that needs to be carried out.

The pricing policy at ELRA and LDC are mainly based on the distinction between research/academic and commercial users. A third criterion at both organizations was based on their own structures, i.e. offering discounted prices to their members.

The prerequisite of acting as a broker is that each purchase renders a payment, covering the compensation claimed by the owner of the resource. In general, as far as ELRA is concerned, ELRA is not the owner of the resources, and can therefore only set a fair price in co-operation with the owner.

In some cases, providers accept to have their resources distributed for free. This is sometimes the case when the production of LRs has already been financed by the European Commission or by national governments. The pricing constraints, sometimes imposed by the providers, are generally of two kinds: a constraint either on the user profile or on the type of use. The providers may limit the distribution to members only, or they may limit the use to research at large or even to academic research only. When the constraints are connected with the type of use, the reason is often that the providers do not want their resource to be used in technical (commercial) development.

Another type of pricing was added recently to ELRA's current offer. ELRA included in its catalogue "Evaluation Packages", which propose not only LRs but also protocols, methodology, tools, etc. that may be used to evaluate Language Technologies. For these Evaluation Packages, a different user license has been created by ELRA in order to limit the use of the packages for the purposes of evaluating Human Language Technologies. Prices are then considered differently depending on whether they are offered for evaluation, research or commercial use.

Finally yet importantly, one should bear in mind that, although the model described herein is based on facts going back to mid nineties, most of this is still valid despite the emergence of the open source trends. One should think of these models also in the framework of Business-to-Business where open source approach is negligible but also in the framework of research that is strengthening the open source/freeware communities. The concept of "sharing" sources that goes beyond open source, freeware to embrace issues like easy licensing, fair market prices, remote use of resources, etc. will be an important part of the discussions of the second FlareNet forum in Barcelona (mid February 2010).

3.6 LRs Maintenance requirements, sustainability model and impact

Very recently, one of the LREC 2008 conference satellite workshops focused on the sustainability issue for Language Resources, highlighting the statement that this issue is regularly neglected in the field of LR production. The workshop was the first one especially devoted to the "sustainability of language resources and tools for Natural Language

Processing" over time. Indeed, for many years now, we have observed that many LRs have been tailored to the needs of specific and individual applications or projects, with no possibility afterwards to access to the LRs developed, no matter the efforts (time, money) needed to create such LRs.

Increased sustainability for linguistic tools and language resources becomes more and more important for the research community. Meanwhile, even funding organizations recognize this fact and the underlying problems – they often encourage research projects to make sure that language resources will be accessible and (re-)usable in 10, 15, or 20 years time but no mechanisms or instruments exist to support such processes over time.

The challenge of ensuring sustainability is a multi-faceted one and depends on several subtasks. Several subtasks addressed in the workshop, out of which:

- Documentation of : content, format, discovery, access, citation, preservation, rights
- Monitoring the evolution of the annotation technology and standardization of annotation frameworks

3.6.1 Need for maintenance of LRs (bug reporting, updates & improvements)

In [Heuvel et al. 2002][1] were highlighted the increasing activities around the validation of resources produced. Nowadays, especially for the speech-oriented resources, a validation procedure has been integrated to the production cycle in order to comply with specifications of production but also with specifications of suitability of resources to the needs of the widest community. Validation work is a first step towards that suitability assessment of resources, but is not the only one.

At ELRA, a bug reporting service was created in order to enable users of ELRA resources to give feedback on the resources they had acquired and with the aim of helping adapt, improve and update those resources.

Other works are also ongoing for the updating of linguistic resources such as lexica. For instance, we can mention the work at the ILC in Pisa on the feasibility to update lexica in a semi-automatic manner in order to diminish the costs of production.

3.6.2 Need for a production/packaging model and a sustainability analysis process

Sustainability is a concept that should be addressed prior to any serious data production. Unfortunately, those who are often in charge of data collection are more interested in the research they could conduct on the basis of that resource than on its potential future. Therefore, it is crucial that sponsors (funding agencies, universities, companies) should require this aspect to be addressed while elaborating the technical specifications and other

[1] Henk van den Heuvel, Khalid Choukri, Harald Höge, *Give me a bug: a framework for a bug report service*, In: LREC 2002 Proceedings, Las Palmas, Spain, 2002.

important required documents. Of course, one can also envisage such concept applied to existing resources and see how one can handle this a posteriori.

Sustainability can be decomposed and hence analyzed along a number of dimensions. Such aspect does not touch upon the LR documentation and LR format only but should go beyond that, as we will describe in a coming deliverable (D2.2). Among these dimension we can list the cataloguing and (hopefully sharing/distribution), the support (in addition to the documentation) of tools to gain access to the data (e.g. visualization, annotation framework), capacity to extend the various annotations and add value over time to the "raw" data but also to report identified bugs and hopefully revise/correct/update/maintain it, keep track of the legal aspect including licensing,

4 Chart of players and classification along different lines

4.1 Hard facts about the HLT Market

This second part aims at illustrating the market status (not necessarily its size & evolution) within the work boundaries we are familiar with. It is hard these days to talk about market size and company structures (SMEs, Larges, Academia, etc.) that are only involved in HLT. We will rather elaborate here on other perspectives that could help us understand the market status.

From the analysis, we will find below, we can see how some areas shrunk very substantially (speech recognition technologies) while new have seen the emergence of new players (MT).

However, let us share some of the recent figures on these two major areas: Speech Technology and Machine Translation. These figures have to be taken with extreme care.

A number of studies and surveys still focus on the market size. For instance, one can read on the Canadian Language Portal (AILIA!) that the "*Global Speech Technology Market to Grow by 36% per Annum, According to New Report By Global Industry Analysts, Inc. (Date Published: February 28th, 2008, Source: eMediawire).*

> *Buoyed by the robust growth in developed regions, global speech technology market is portended to reach US$7.8 billion by 2010. North America and Europe collectively contributes to over 91% of expenditure on speech technologies. Automatic speech recognition market represents the largest segment. Sales from automatic speech recognition products are expected to register about US$7.5 billion by 2010.*

Even if the current crisis has an impact on these figures, they remain very impressive though hard to confirm through other channels. We have also considered the survey conducted by VanDijk for the French ministry of Research & New Technologies (in the framework of the Technolangue programme) which has a very limited scope and aims to assess the European market as regards the language tools that process speech and text, and identifies the axes of development. The figures are still available but probably hard to interpret and somehow obsolete (http://www.technolangue.net/IMG/pdf/SyntheseUKEtudeMarche-Technolangue2006.pdf)

If we jump to the market of Translation, and in particular of Machine Translation, players like Language Weaver anticipate a market of over US$67 billion of what they refer to as "Digital Translation Market". This is based on the digital content being produced and assuming 0.00001% would require some translation and would boost the demand thanks to the "low cost of quality translation". Language Weaver CEO Mark Taplingindicated that "Conventional language translation services are typically priced at 21 cents per word, leaving many high volume requirements behind. Language Weaver is able to provide an automated solution that dependably conveys communication meaning for applications with high volumes, requiring speed, and accuracy". Common Sense Advisory search firm indicated that many big players (in areas like Web Content, Business Intelligence, and Customer Care) would go for MT, as "automated translation is a way to solve logarithmic growth in content volume, velocity, and volatility".

To compare with the figures given by Sheryl Hinkkanen, FIT Secretary General (extract from a survey conducted by Allied Business Intelligence, Inc. USA): The machine translation market also increased robustly, from 324 million US dollars in 1999 to 447 million US dollars in 2004, a rise of nearly 40%.

4.2 Different dimensions to describe players' profiles

As indicated all through this document, the chart of players within the HLT sector aims at sharing with the readers some views on the non-purely commercial indicators that may describe the market. It should be taken as quantitative descriptors rather than turnovers, profits, and the like.

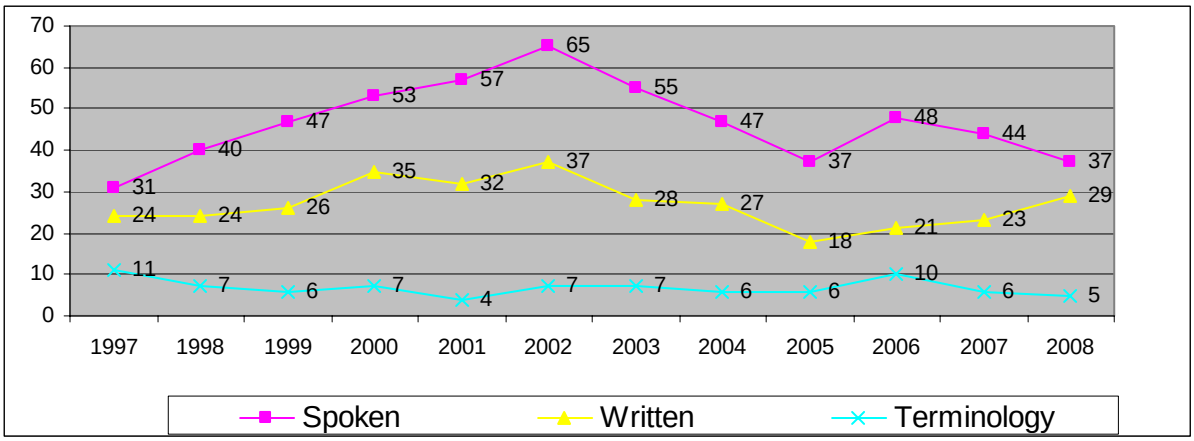
The figures that will be displayed in this section come from different databases maintained by ELRA and other partners and which aim at identifying the players of the HLT field. In particular, those players are people either who participated in the LREC conference organized by ELRA (and which gathers about 1,000 participants), or who have been listed by ELRA for various reasons: they can be members of ELRA, providers, customers or receivers of ELRA's publications, they could be related to ISCA, EAMT, LDC, etc.

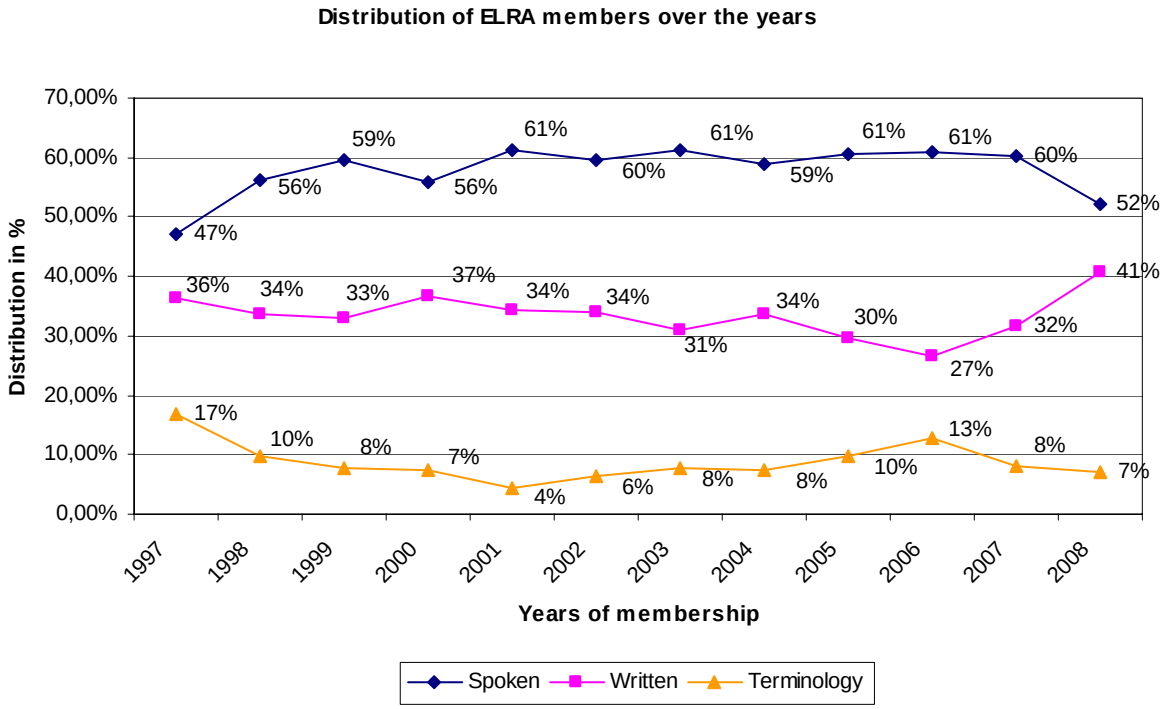
Thanks to those lists, several dimensions can be highlighted with respect to the following criteria:

- **field of interest (speech, written or terminology)**
- **region of location**
- **type of organizations (e.g. company versus academia)**
- **market**

4.3 Description of players per field of interest

ELRA maintains statistics regarding the evolution of its members over the years. If we compare those statistics with the sales of language resources, we can easily highlight the evolution in terms of field of interest (speech, written or terminology area).

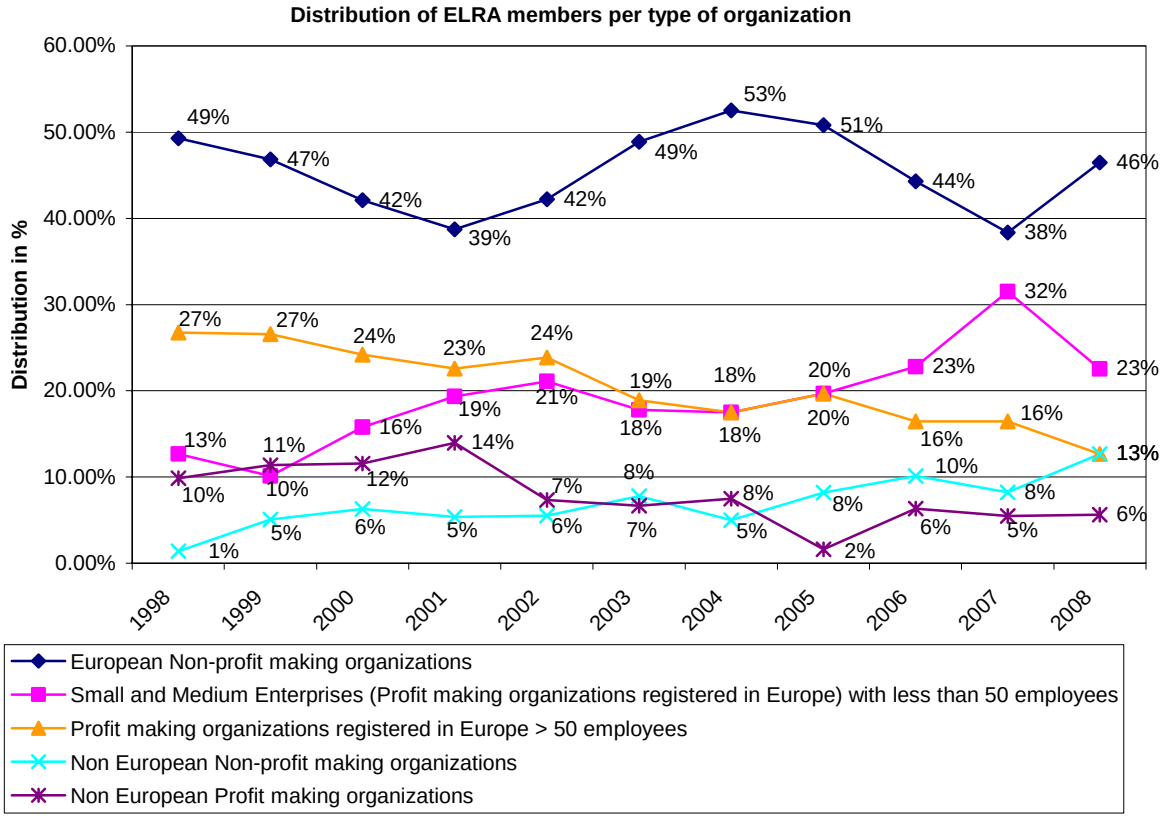




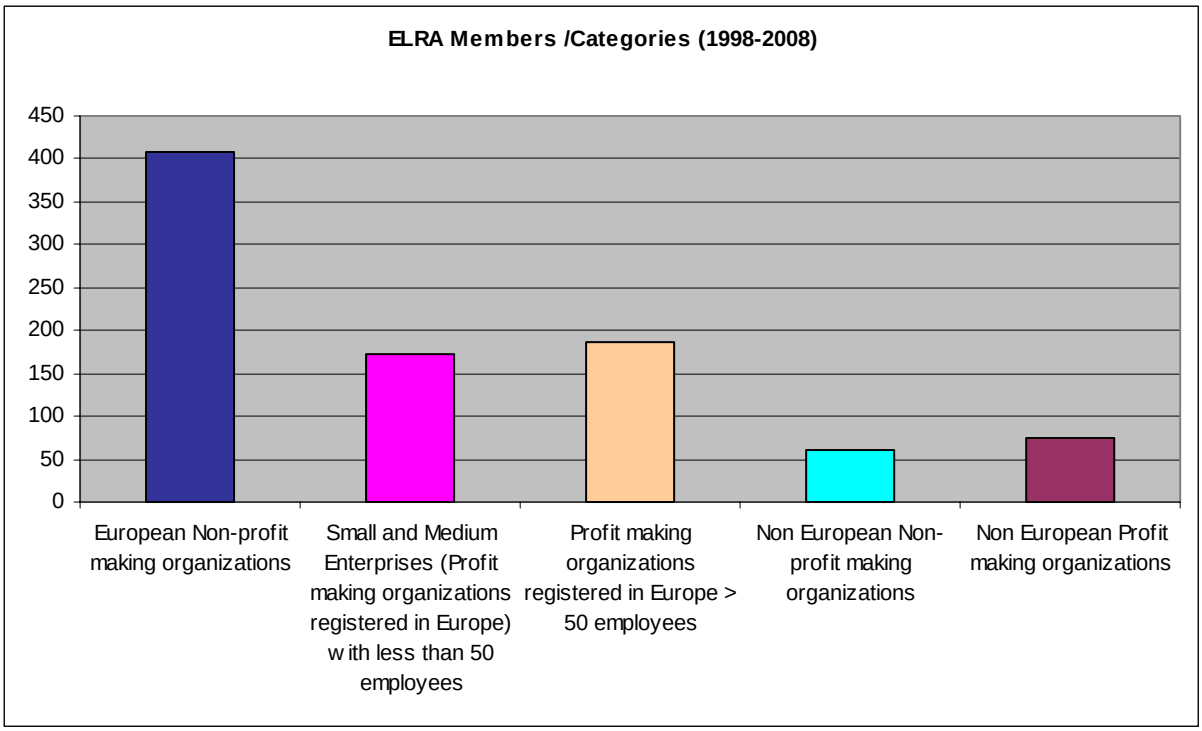
We see from these diagrams the importance of speech technology deployment in the HLT area versus the other domains. It seems that ASR players are more than other prepared to share data (under some market conditions). This suggests that in this domain, the added value is not solely in the data (though all players use data driven approaches): the specifications and best practices are well know and major players can easily afford to produce the data they need. So Producers (often technology developers) prefer to offer the data for sale at a partial production costs to recoup their investment even if this allows competitors to gain a few months that would be required by the production. This is not the case for other areas (that for instance required large lexicographical data) but this is changing with the strength of the web.

4.4 Description of players per type of organization

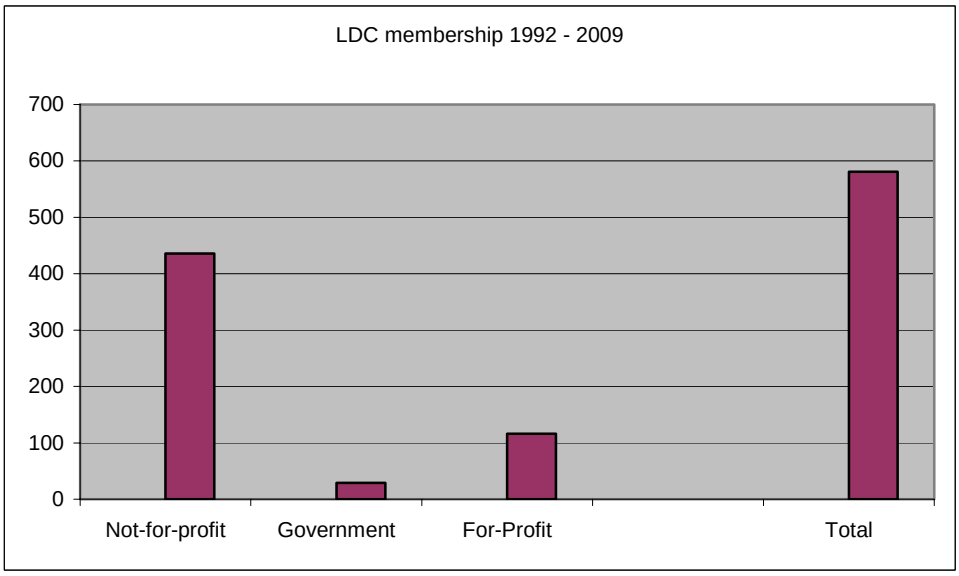
Below is given a graph showing the distribution of ELRA members with respect to their type of organizations:



A cumulated diagram is given herein:

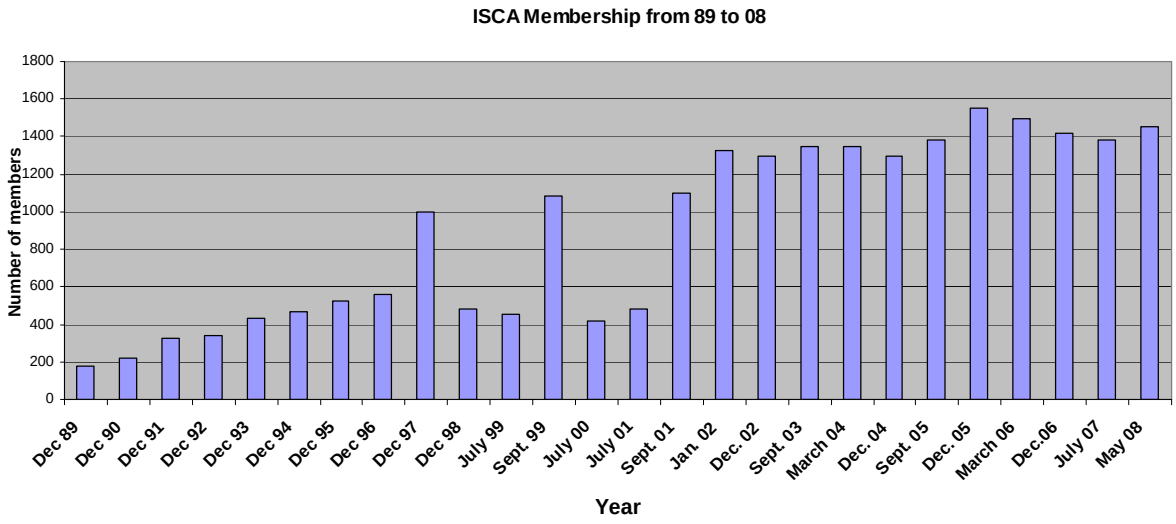


A similar concise diagram illustrates the LDC membership cumulated between 1992 and 2009:

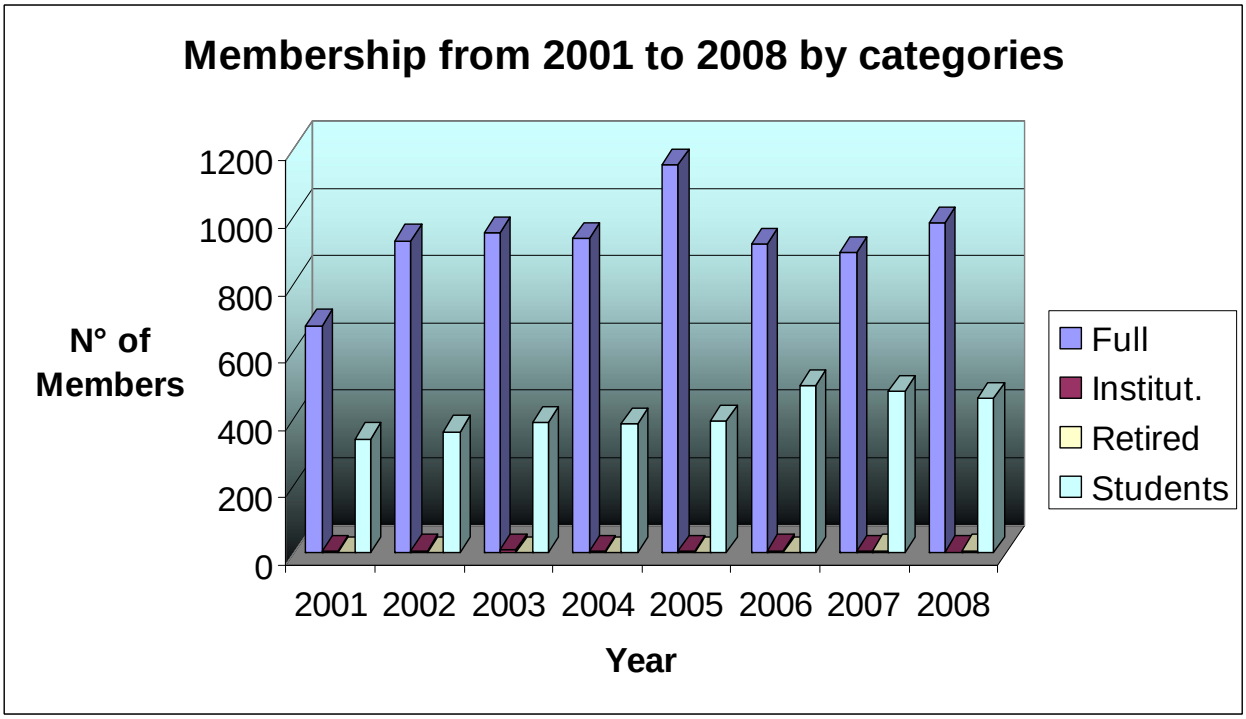


If we focus on the particular area of speech technologies and specifically through the facts and figures provided by ISCA (International Speech Communication Association), the major organization representing the speech community, we can see that, despite all consolidation

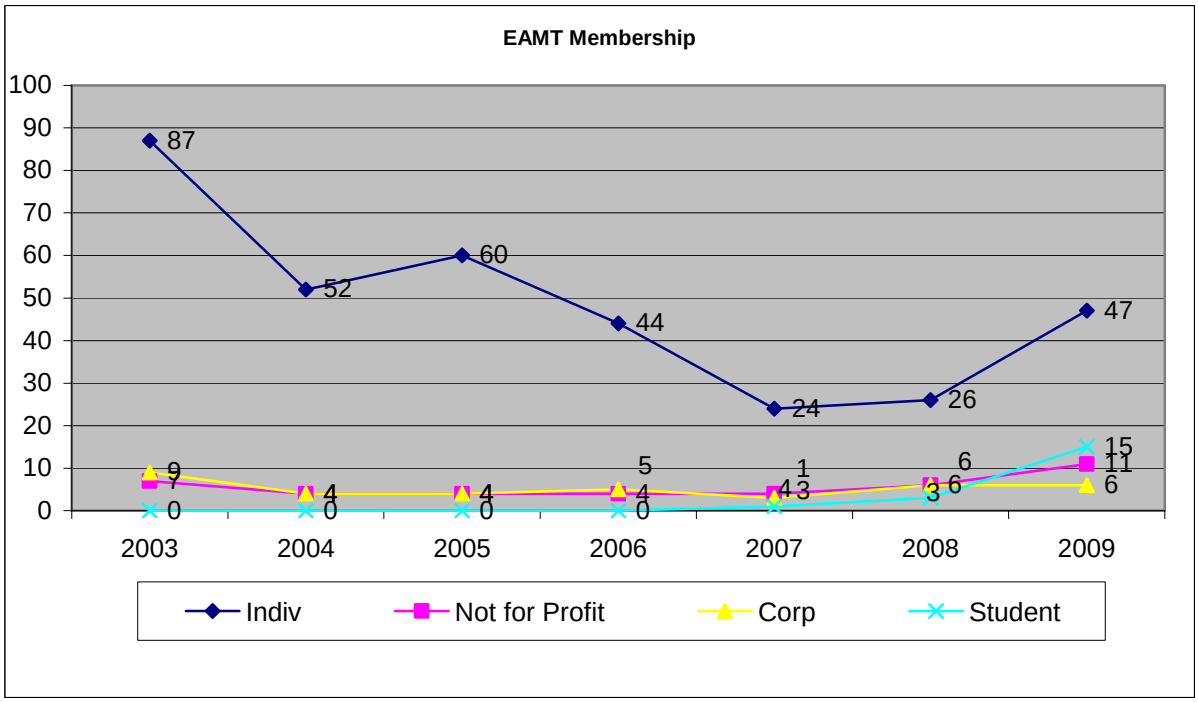
seen in the industry, the research activities are still very strong, given the number of ISCA members (also participants of Interspeech conferences, up to 2008):



Since the ISCA membership comprises individuals (both students and staff members as well as institutions), the following diagram highlights how this is distributed:



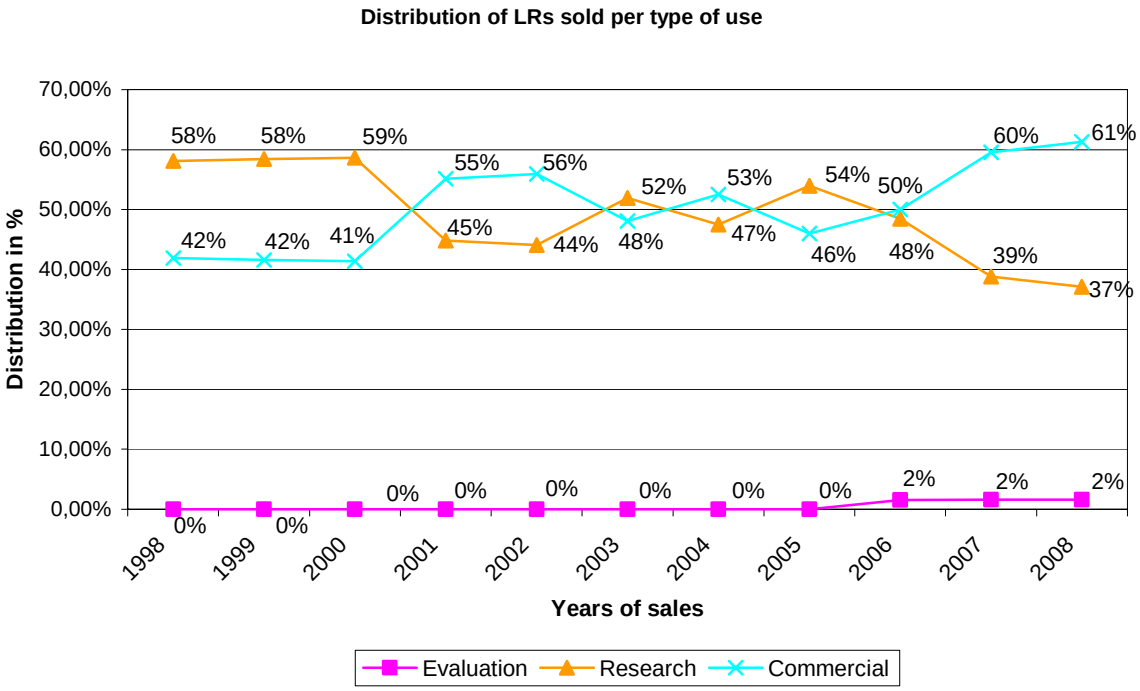
For EAMT, the following membership illustrates the involvement of the community in MT:



From the data gathered by ELRA, LDC, ISCA, EAMT we see clearly that the traded resources are of interest to the non-for profit players (mostly R&D labs under public organizations)

4.5 Description of players per type of usage (R&D versus commercial)

This another dimension that is illustrated below for the descriptions of players: the type of use they make of the LRs, which is closely linked to their type of organizations. For example according to ELRA's licensing policy, research use applies, most of the time, to academic organizations whereas commercial use applies to commercial organizations. Therefore, the graph well highlights the increase of commercial organizations interested in acquiring LRs versus a small decrease of interest from academic organizations. We can see the emergence of organizations interested by evaluating their technologies through the acquisition of evaluation packages. Here, however, we cannot extract a differentiation between academics and commercial organization, although from a quick expertise, we could easily presume a high interest by commercial organization to exploit evaluation packages to improve the performance of their tools. ELRA has introduced a third category that is called evaluation covering the acquisition of evaluation packages to assess technology performance. Such us is exclusive of any research or development activity and therefore was set aside on purpose. The data from LDC does not reflect such issue as LDC distributes the data for R&D only.

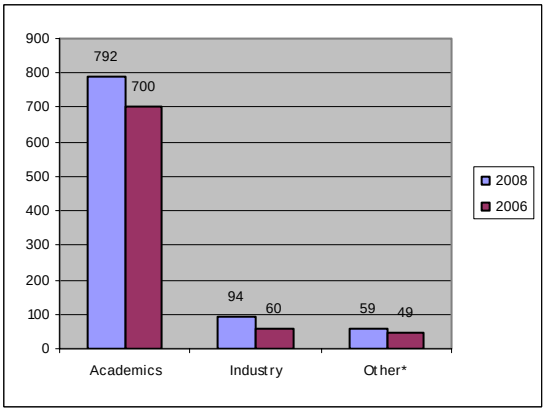


Over a long period, it seems that ELRA serves both communities and this is an important input for FlareNet recommendations to better sponsor the research activities.

If we consider the data from LREC regarding the profile of participants, we can see data that confirms the distribution figures. Revisiting LREC 2006 and 2008 figures, we can draw some quick conclusions about the profiles of the participants:

	2008		2006	
Academics	792	**84%**	700	**87%**
Industry	94	**10%**	60	**7%**
Others*	59	**6%**	49	**6%**

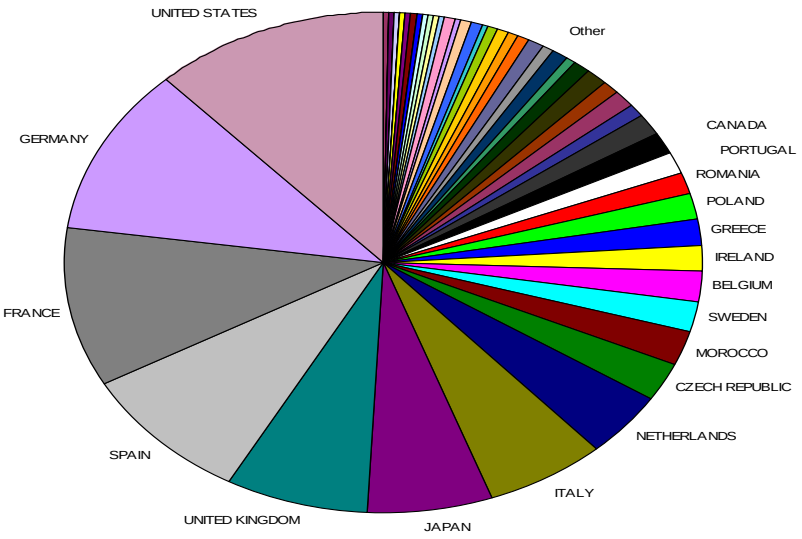
*Others include public institutions such as US National Library of Medicine, Basque Government, NIST, the European Commission,



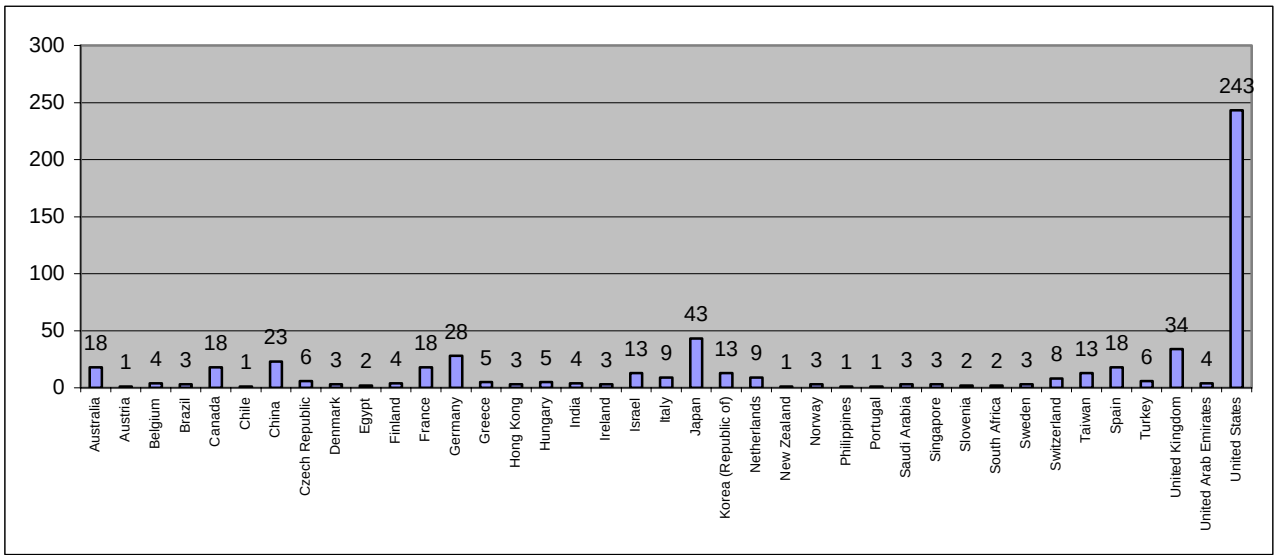
4.5.1 Description of players per country of origin

In addition to the diagram of section 4.4, we can look at LREC 2008 participants' data from which we can retain another dimension: the country of origin of the organizations. In the graph below, we can easily see the high number of players coming from Europe, with a main participation from organizations from France, Germany, Italy, Spain and United Kingdom.

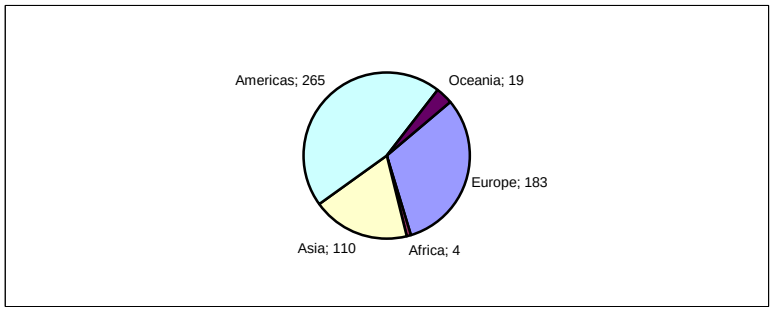
Then, comparatively poor is the number of countries from Eastern Europe, which can be considered as a hint for future developments.



For LDC cumulated membership, we have:

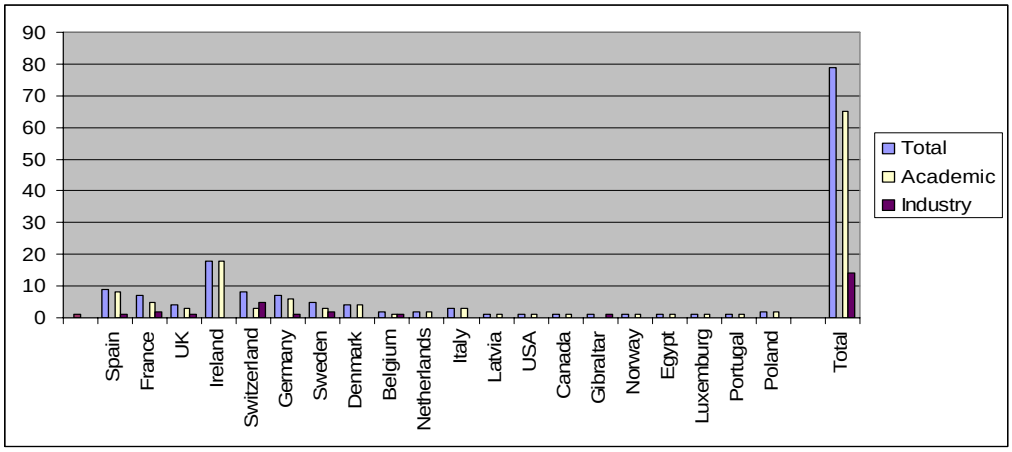


or per continent:

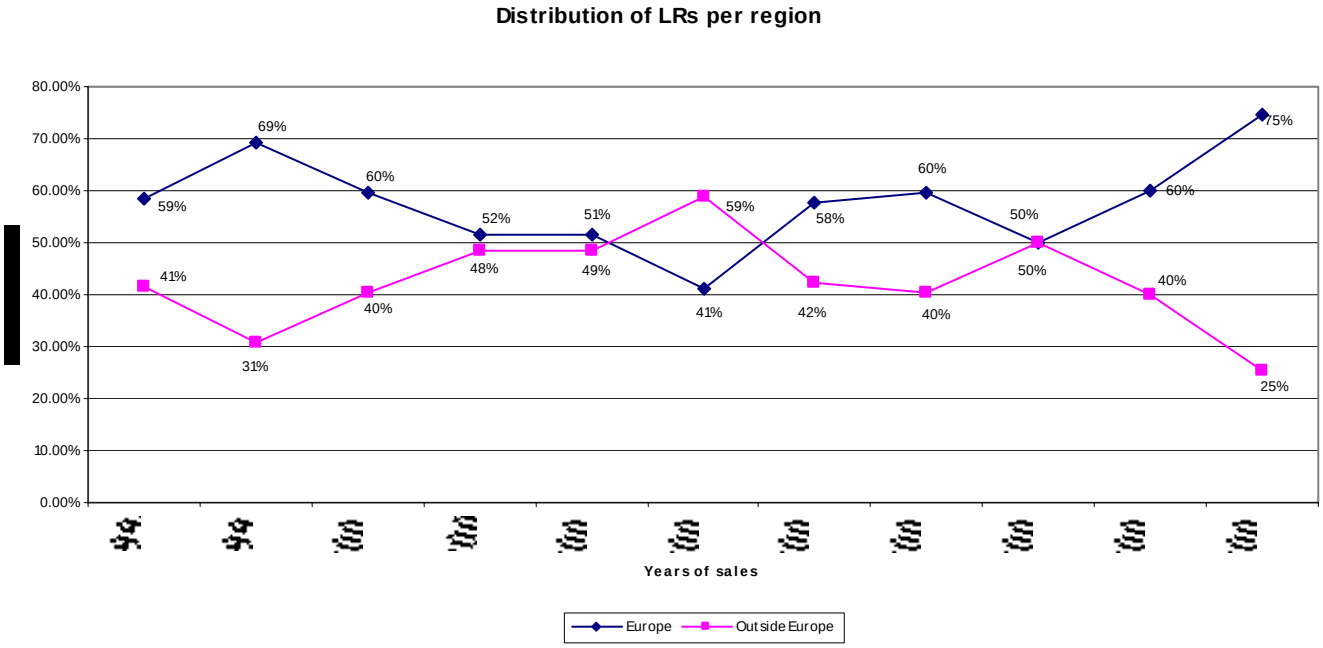


If we consider the ELRA market of LRs with respect to the location of the organizations who acquired LRs, we can see that European organizations still remain in highest demand of ELRA LRs. This can subsume most probably that the market outside Europe can be answered by providers outside Europe instead of European organizations such as ELRA.

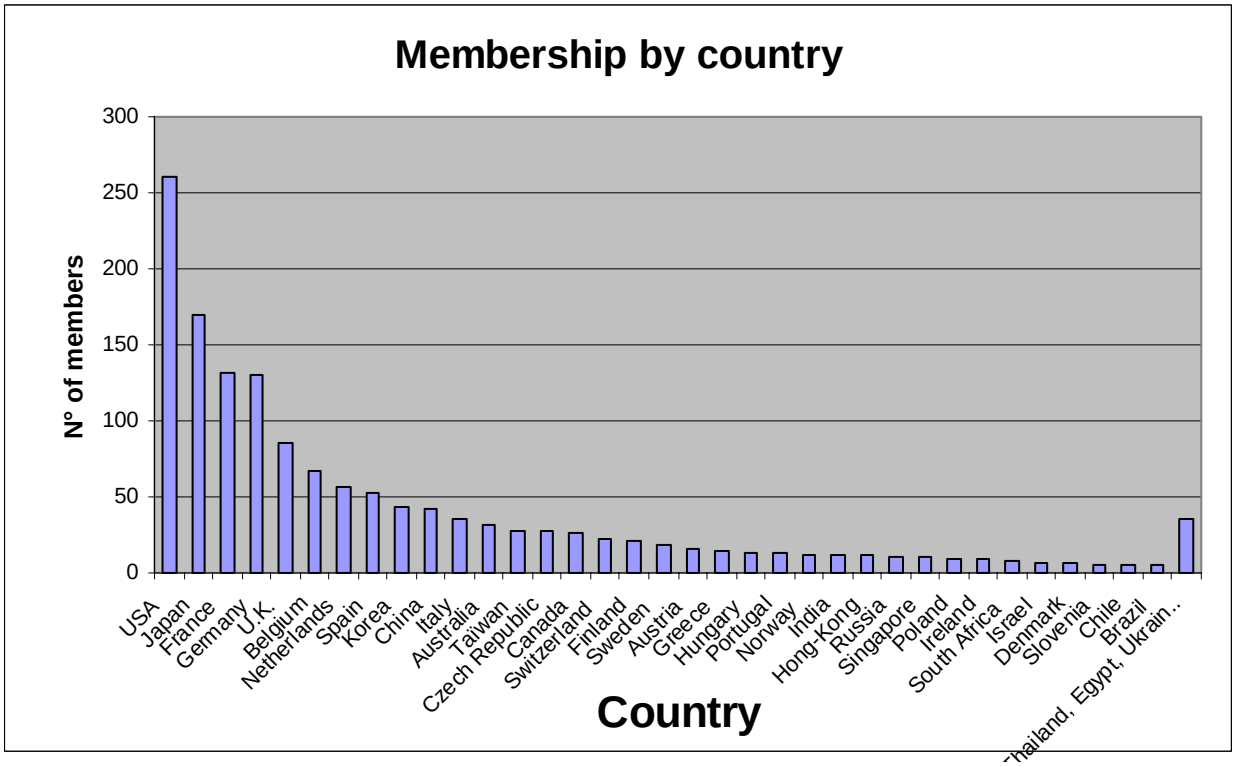
The members of EAMT (EAMT focuses on Europe and the Middle-east since this year) show the following countries represented within its membership (non-Europeans could also join):



We can also see what the different players acquire from ELRA in terms of Language Resources with respect to their location:

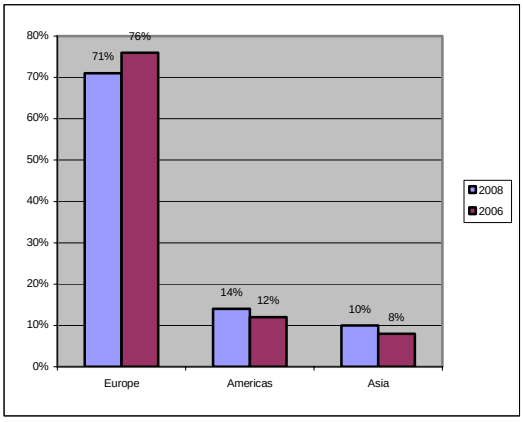


We can also see this at the scientific level through the ISCA membership distribution that comforts the previous findings:



In addition to that, we can also see this geographical distribution through the number of participants to LRECs per country, a good indicator of both active R&D and industry in the field.

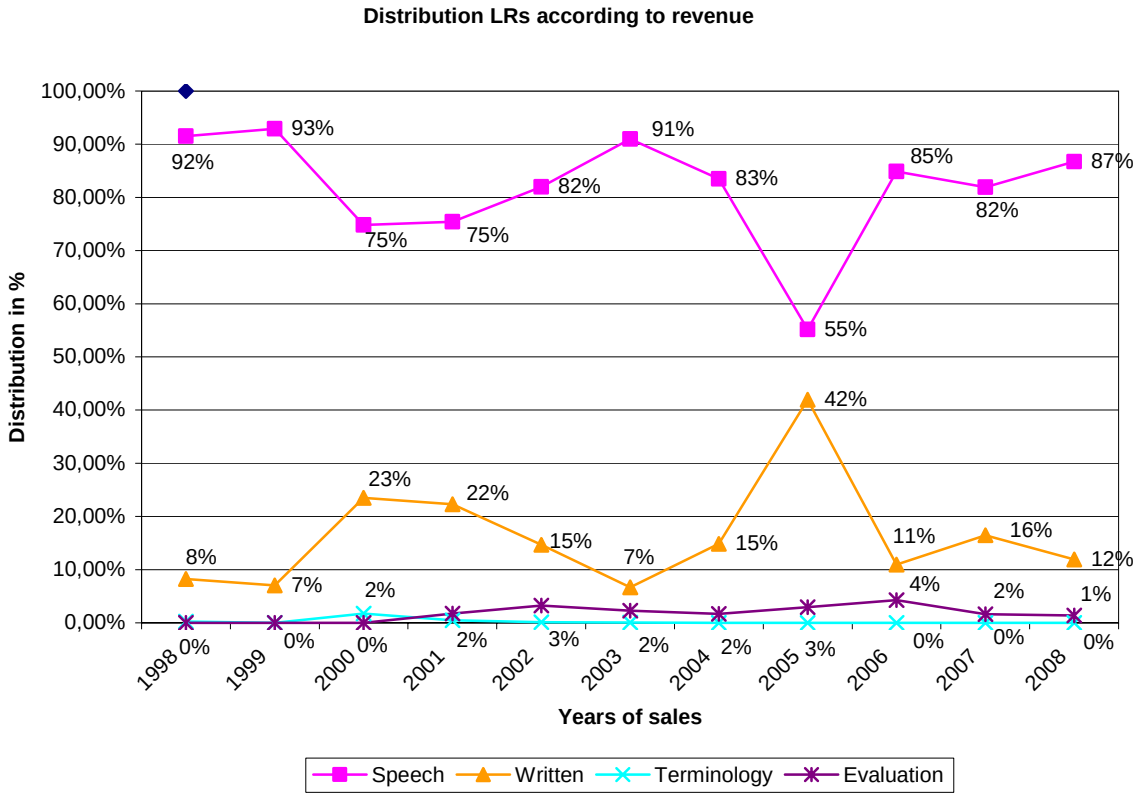
		2008 (Marrakech) 1100 participants *57 countries*		2006 (Genoa) 800 participants *44 countries*	
Top countries	**France**	3	**10%**	4	**9%**
	Germany	2	**11%**	1	**12%**
	Italy	7	**6%**	3	**11%**
	Japan	6	**6%**	6	**6%**
	Spain	4	**9%**	7	**6%**
	UK	5	**7%**	5	**7%**
	USA	1	**12%**	2	**11%**
Continents	**Europe**	1	**71%**	1	**76%**
	Americas	2	**14%**	2	**12%**
	Asia	3	**10%**	3	**8%**



Through these figures, the "big" countries (France, Germany, UK, Italy, Spain,) show a stronger involvement in the HLT activities.

4.5.2 Other Market dimensions

Beyond the dimensions mentioned above, another dimension is also relevant to show the market capacity of those players. Indeed, ELRA and its distribution agency regularly analyze the distribution of players according to the revenues obtained from the sales of LRs. Such information if compared to the previous tables can show some discrepancies. In particular, if we compare the interest of players with respect to types of resources to the revenues obtained, we can see that revenues are higher for Speech resources than for Written resources, although Written resources seem to be more used (and thus needed).



5 Conclusions and next Phase plans…

5.1 Chartering LRs within FlareNet … which perspectives ?

The first part of this report aimed at giving the reader a brief description of the Language Resources of use within the Human Language Technology. It elaborates on potential features and dimensions to describe them. While the report did not claim to be another inventory of existing LRs, it gave some data about existing catalogues (designed for HLT).

The most critical analysis one can draw from this is the large number of enthusiast initiatives that collect information on LR and describe them. It is a pity that the only project that aimed at harmonizing such initiatives is OLAC, which is based on selecting a minimal set of metadata elements (very small regarding the HLT needs) for all types of LRs. FlareNet has planned that in Phase 2 it will strongly support tasks it initiated with ELRA and LDC to boost such harmonization and even go beyond that in supplying a uniform and coherent description and search frameworks.

Another important task fro phase 2 is the monitoring of trends and evolutions of LRs (as briefly introduced in this report). A number of key resources will be identified (about 20 to 50), covering all HLT areas and will be described along a number of critical dimensions over time (e.g. languages, sizes, design and specs, other technical characteristics …). The idea is to see how such resources depict the efforts devoted by the community to maintaining, improving, updating, etc. Language Resources. FlareNet and ELRA have already launched this task through the project “LREC Map” that aims to collect information about all resources mentioned in LREC submissions (used, produced, updated, etc. by the authors) and other initiatives will be launched with LDC.

This first part also tackled the issue of BLARK for a given language, arguing for the design of a BLARK for a given language and technology with strong coordination at the technical and political levels. It attempted to draw a better picture of the minimal size of data required for training baselines, taking into consideration the compromise between performance to achieve, quantity of data, and its cost.

The report indicated also that the crucial issue of self-sustainability will be elaborated upon on another report.

5.2 Chartering HLT Players within FlareNet … which perspectives ?

For the part devoted to HLT market and players, we tried to stress that players were our main focus and not the market: we wanted to describe their profiles, sectors of activities, technologies, etc. instead of reproducing the usual market values and turnovers that are hard to judge, particularly in a crisis periods; The first fact one should consider is the strong consolidation that have been executed in several areas and the shrinking of the number of key players. For instance, though the facts are based on ELRA data only, it is clear that commercial organizations within the speech sector are less numerous than by the past. On the contrary, a number of commercial players emerged in other sectors like MT (or SMT) and became more visible thanks to web-based translation services (mostly offered for free to the

general public) and also, and this is very new, to the technological evaluation campaigns conducted worldwide (NIST MT evals, ELDA TC-STAR and CESTA, IWSLT, etc.).

Another remark from these figures is the unbalanced ratio between research and commercial usage of resources. Although most of the LR revenues (for data centers, distributors, right holders, etc.) is generated by commercial organizations, the number of resources supplied to researchers is, by large, more important. It is clear that the non-traded resources are exchanged bi-laterally between players under confidential agreements.

Another remark is about the geographical locations of the players. Though R&D events (e.g. conferences, evaluation campaigns, etc.) attract a large number of Europeans, it is clear that more activities are located in the USA and Japan. This remark should be taken with care given the globalization process in place at the major players with the outsourcing and relocation of R&D labs (Orange Labs, Microsoft, IBM, Google, etc. to Egypt, China).

Last but not least, despite the consolidation/shrinking mentioned above, the reduction in R&D budgets, it seems that major conferences (LREC, Interspeech, MT summit, ICASSP) attract more and more scientists.