# The LREC 2010 Resource Map

[1]Nicoletta Calzolari, [1]Claudia Soria, [1]Riccardo Del Gratta,
[1]Sara Goggi, [1]Valeria Quochi, [1]Irene Russo,
[2]Khalid Choukri, [3]Joseph Mariani, [4]Stelios Piperidis

[1]CNR-ILC "A. Zampolli", [2]ELDA/ELRA, [3]LIMSI/IMMI-CNRS, [4]ILSP "Athena" R.C.
[1]Pisa - Italy, [2,3]Paris - France, [4]Athens - Greece
nicoletta.calzolari@ilc.cnr.it, claudia.soria@ilc.cnr.it, riccardo.delgratta@ilc.cnr.it,
sara.goggi@ilc.cnr.it, valeria.quochi@ilc.cnr.it, irene.russo@ilc.cnr.it,
choukri@elda.org, Joseph.Mariani@limsi.fr, spip@ilsp.gr

**Abstract**

In this paper we present the LREC Map of Language Resources (data and tools), an innovative feature introduced in conjunction with the LREC 2010 Conference. The purpose of the Map is to shed light on the vast amount of resources that represent the background of the research presented at LREC, in the attempt to fill in a gap in the community knowledge about the resources that are used or created worldwide. It also aims at a change of culture in the field, actively engaging each researcher in the documentation task about resources. The Map has been developed on the basis of the information provided by LREC authors during the submission of papers to the LREC 2010 conference and the LREC workshops, and contains information about almost 2000 resources. The paper illustrates the motivation behind this initiative, its main characteristics, its relevance and future impact in the field, the metadata used to describe the resources, and finally presents some of the most relevant findings.

## 1. Why a Map of Language Resources?

The purpose of this paper is to introduce the LREC Map of Language Resources (data and tools), an entirely new instrument that has been developed in the framework of the LREC2010 conference[1]. The term "map" suggests a representation of the salient characteristics of a given territory, thus enabling the knowledge and discovery of its main features. A map is drawn in order to make new territories known, or to improve the knowledge of already discovered ones. Why should the "territory" of Language Resources need a map? Several institutions worldwide maintain catalogues of language resources (ELRA[2], LDC[3], National Institute of Information and Communications Technology (NICT) Universal Catalogue[4], ACL Data and Code Repository[5], OLAC[6], LT World[7], etc). However, it has been estimated that only 10% of existing resources are known, either through distribution catalogues or via direct publicity by providers (web sites and the like). The rest remains hidden, the only occasions where it briefly emerges being when a resource is presented in the context of a research paper or report at some conference. Even in this case, nevertheless, it might be that a resource remains in the background simply because the focus of the research is not on the resource per se.

Knowledge about existing resources is essential to the overall advancement of research in the field: it is important to be able to locate and retrieve the right resources for the right applications, and to exploit existing ones before building new ones from scratch. Having a clear picture of which resources are available for which languages and for which use is important in order to identify existing gaps for certain languages at a given time and estimate the amount of investment needed to fill them in.

Knowledge about the current use of resources is equally important. Knowing which resources are most used for the various applications will help to better understand the reason behind their success (their intrinsic quality, their wide availability, their licensing model, etc.). Knowing which standards are used in resource representation would help improve the development of standards themselves, by getting them more tuned to actual needs and requirements. Clear and easy-to-reach information of this type about resources and related technologies is lacking. At the same time, it is very important to stress that most resources are very poorly documented, or not documented at all, thus hindering their accessibility and in the end, their full deployment.

We decided to exploit the unique opportunity offered by the LREC conference of gathering all major players of the sector in order to discover the resources directly or indirectly connected with the research presented at the conference. This felicitous conjunction of people and resources, we believe, will yield an unprecedented and comprehensive overview of the language resources currently being developed and used.

## 2. Drawing the Map

In order to elicit the information needed to draw the Map of Language Resources, it was decided to couple the request of information about resources with the paper submission procedure. This allowed to maximize the amount of information that could be derived by reducing to a minimum the

[1]This work was partially funded by the FLaReNet Thematic Network (http://www.flarenet.eu).

[2]http://catalog.elra.info/

[3]http://www.ldc.upenn.edu/Catalog/

[4]http://facet.shachi.org/?ln=en

[5]http://www.aclweb.org

[6]http://www.language-archives.org/

[7]http://www.lt-world.org/

burden on people entering the information.

### 2.1. Metadata

Since we did not want to overburden authors during the submission procedure, we aimed at a simple form for describing the resources used or created in ones own research. Consequently, the metadata fields and values have been intentionally oversimplified.
The form consists of 12 main metadata fields, which should provide the minimum set of relevant and useful information about new and existing resources and their use.
A set of 9 first metadata has been revised after the abstract submission phase in order to slightly increase the descriptive parameters requested upon submission of the final papers, while an additional set of 3 were added in the final submission phase.

**The first set of metadata** These fields represent quite general information available in most LR catalogues (such as LDC and ELRA) and surveys (e.g. ENABLER[8]). Each of these basic fields has a list of suggested values, which has been deliberately kept short by using only most frequent and common values. However, the possibility has been left open for the user to select an "Other" field and to specify a more appropriate term in case he/she did not feel any of the suggested values were satisfying. This set contains:

- Resource Type
- Resource Name
- Resource Production Status
- Use of the Resource
- Language(s)
- Modality
- Resource Availability
- Resource URL (if available)
- Resource Description

**Additional set** Three descriptors ("Resource Size", "Resource License" and "Resource Documentation") were added in the final submission phase to allow for extraction of additional information.

### 2.2. The Tool

The tool designed for entering information about language resources was integrated into the START submission page so that authors could provide their information during the standard submission procedure. At the same time, this allowed an efficient link among papers submitted, related resources and authors of the paper. The graphic aspect of the tool was kept deliberately simple: for each resource that was either described in the paper or had been used for the research to be reported about, authors had to fill in a form containing as many fields as the metadata described above. Up to ten resources could be inserted for each paper submitted.
A more complex graphic version of the tool will also be available on the web. Its functionalities are the same, but it can be easily extended to provide new features. This web tool has been designed for managing different conferences so that it will be possible to easily make comparisons among the information provided at the various conferences. In addition, the web tool offers the authors the facility of adding new resources on demand, independently of a specific conference/event. Finally, the tool will be integrated with a search engine for browsing information.

[8] http://www.ilsp.gr/enabler/

## 3. Reading the Map

Response to the Map was generally very good, with a total of 1994 entries provided. The impressive amount of information gathered holds an enormous potential in terms of analysis that can be performed.
The great deal of data provided by the Resource Map can be analyzed according to different dimensions of analysis. Some dimensions are directly linked to the metadata described in 2.1., while others can be extracted from information contained in the general START database.
The starting point of our analysis is based on the metadata used to fill in the form: the analysis can be performed either mono-dimensionally, i.e. by taking into account one metadata element at a time, or multi-dimensionally, by combining two or more metadata elements and looking for the various correlations.

### 3.1. Monodimensional Analysis

Analysis along single metadata elements allows to extract coarse-grained information which is only partially similar to that already available in current catalogues. The "ResourceType" descriptor helps to overview the different typologies of language resources, and will very likely complement the range of information already available elsewhere. Similarly, the "Use of the Resource" descriptor helps us assess the distribution of the resources according to their uses, a piece of information that can give interesting results when compared with similar data already surveyed by other catalogues. On the other hand, the analysis of the results along the dimension provided by the "Resource production status" element gives us an entirely new insight about the number of newly created resources versus existing/used ones, and on the extent to which some resources are used more than others, their particular type and languages, etc.

#### 3.1.1. Resource Type

The "Resource Type" descriptor was aimed at capturing the general type of the resource used or created being described. A wide notion of resource was adopted that encompasses not only the usual types of resources, such as corpora and lexicons, but also tools, metadata, guidelines and standards, evaluation data, tools and methodologies, were considered as resources.
Corpora appear to be the most frequent type of resources, with 785 instances in total. Table 1 reports the values for those types recurring more than 20 times in the whole Map. It is also possible to derive, for each category of resources, the one that is mostly used across several applications and researches. For instance, it appears that the most frequently

| Type | No. of instances |
|---|---|
| Corpus | 785 |
| Lexicon | 239 |
| Tagger/Parser | 181 |
| Annotation Tool | 134 |
| Ontology | 73 |
| Evaluation Data | 40 |
| Representation-Annotation Formalism/Guidelines | 39 |
| Grammar/Language Model | 32 |
| Evaluation Tool | 32 |
| Terminology | 29 |
| Named Entity Recogniser | 29 |
| Representation-Annotation Standard/Best Practice | 23 |

Table 1: Most frequent types of resources

used Corpus is the World Wide English Corpus, followed by Europarl. In the "Lexicon" category, the most widely used one is WordNet, followed by FrameNet. See Table 2 for more details.

| Resource Type | Name | Citations |
|---|---|---|
| Corpora | World Wide English corpus | 20 |
| | EUROPARL | 14 |
| | Wikipedia | 11 |
| | British National Corpus | 7 |
| | National Corpus of Polish | 7 |
| | Prague Dependency Treebank 2.0 | 6 |
| | PAROLE | 4 |
| | Penn Arabic Treebank 2 | 4 |
| | Prague Czech-English Dependency Treebank | 4 |
| | SoNaR | 4 |
| | The GENIA corpus | 4 |
| | ... | ... |
| Lexicons | WordNet | 10 |
| | FrameNet | 5 |
| | The EDR Electronic Dictionary | 4 |
| | EuroWordNet | 3 |
| | General Inquirer | 3 |
| | Hindi Wordnet | 3 |
| | Lefff | 3 |
| | PDT-VALLEX | 3 |
| | SentiWordNet | 3 |
| | DIINAR.1 | 2 |
| | ItalWordNet | 2 |
| | ... | ... |

Table 2: Most frequent Corpora and Lexicons

#### 3.1.2. Resource Production Status

The purpose of this descriptor was to discover whether a given resource presented in a paper already existed or was an entirely new one. For a newly created resource, we wanted to know whether the resource production was completed or if work was still in progress. In the case of an existing resource, we were interested in discovering whether it had been simply used ("Existing-used") or else updated or modified ("Existing-updated").

From the point of view of the production status of the resources, i.e. whether a resource is new or is being used, it appears that the field is very active with a wide proportion (44%) of new resources being created. Existing resources amount to the remaining 56%. Among these, the majority (83%) is just used, while only a minor portion of existing resources (17%) is being updated in some way. On the other side of the new resources, only 32% are finished, while the remaining 68% are still being developed.

The "Resource Production Status" parameter is particularly interesting when combined with others as it yields a view on which types of resources are mostly used or else, about the kind of resources that are newly built (see below, Section 3.2.).

#### 3.1.3. Languages

The LREC Map registers resources for 170 different languages, with an obvious prevalence of English. A list of the ten most cited languages is given in Table 3.

Monolingual resources constitute the vast majority, as illustrated by Table 4 below.

#### 3.1.4. Modality

For this parameter the suggested options were "Written", "Speech", "Multimodal/Multimedia", and "Sign Language". An option "Other" could also be chosen in case none of the above descriptors applied. "Not applicable" was the choice for those types of resources for which modality is not a relevant descriptive parameter, such as tools or ontologies, for instance. Table 5 below illustrates the results.

#### 3.1.5. Use

The purpose of this descriptor is to make the "actual use" of the resource emerge. For an existing resource, we wanted to know for which application/task the resource was actually used in a given research. For any newly created resource it was asked to indicate the actual/intended use for which the

| Language | Citations |
|---|---|
| English | 723 |
| French | 187 |
| German | 166 |
| Spanish | 127 |
| Italian | 118 |
| Arabic | 74 |
| Dutch | 74 |
| Japanese | 72 |
| Chinese | 68 |
| Portuguese | 52 |
| Swedish | 50 |
| Czech | 48 |
| Greek | 34 |
| Romanian | 33 |
| Basque | 32 |
| Catalan | 30 |
| Polish | 28 |
| Danish | 25 |
| Hindi | 23 |
| Hungarian | 22 |

Table 3: The 20 most cited languages

| Type | % |
|---|---|
| Monolingual | 73 |
| Bilingual | 15 |
| Multilingual | 7 |
| Trilingual | 5 |

Table 4: Monolingual vs. multilingual resources

resource has been built.
Replies were extremely varied. A striking finding is the high proportion of user-defined descriptors: 53% of tags were provided by users. Figure 1 illustrates the top-5 categories of most frequent uses specified for the whole amount of resources. A longer list is available in Table 6.

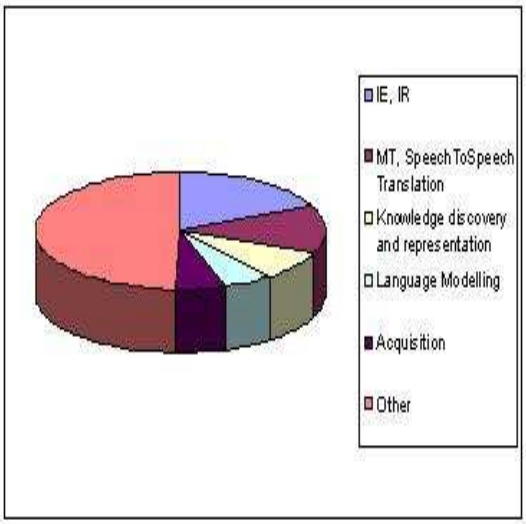


Figure 1: Most frequent uses of resources

| Modality | No. of instances | % |
|---|---|---|
| Written | 1509 | 79 |
| Multimodal/Multimedia | 163 | 9 |
| Speech | 133 | 7 |
| Other | 59 | 3 |
| Sign Language | 45 | 2 |

Table 5: Modality values

| Application | % |
|---|---|
| Information Extraction, Information Retrieval | 16 |
| Machine Translation, SpeechToSpeech Translation | 12 |
| Knowledge discovery and representation | 7 |
| Language Modelling | 5 |
| Acquisition | 5 |
| Emotion Recognition | 4 |
| Discourse | 4 |
| Named Entity Recognition | 4 |
| Word Sense Disambiguation | 4 |
| Question Answering | 3 |
| Speech Recognition | 3 |
| Dialogue | 3 |
| Document Classification, Text categorisation | 3 |
| Web services | 3 |
| Semantic Web | 3 |
| Language Identification | 2 |
| Text Mining | 2 |
| Part-of-Speech Tagging | 2 |
| Sign Language Recognition/Generation | 2 |
| Textual Entailment | 1 |
| Parsing | 1 |
| Speech Synthesis | 1 |
| Natural Language Generation | 1 |
| Summarisation | 1 |

Table 6: Most frequent uses for all resources

#### 3.1.6. Availability

This parameter allows to highlight the different means by which resources are distributed and in particular, to assess the extent to which resources are freely available for community use. Possible choices were:

- Freely available: resources/tools available on the web, at least for research
- From Data Centers: e.g. ELRA, LDC,...
- From Owner: resource distributed directly by the owner
- Other: any other option if needed

Figure 2 below represents the proportion of the different categories with reference with the Availability Status. See also Table 7 for more details. The wide majority of resources (54%) are freely available, while a significant proportion (28%) is only available directly from the owner.

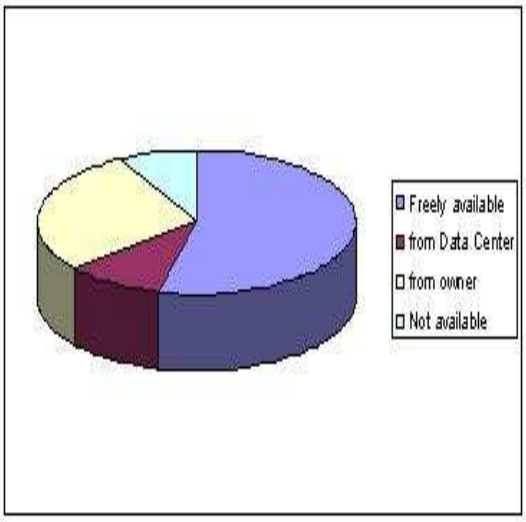


Figure 2: Availability status

| Availability | % |
|---|---|
| Freely available | 54 |
| From Owner | 28 |
| From Data Center | 10 |
| Not Available | 8 |

Table 7: Availability status for all resources

#### 3.1.7. Type of Licence, Documentation and Size

These descriptors were added in the second phase of the Map elicitation procedure. They are significantly less populated as the majority of authors did not review their resource description. As far as the Type of Licence is concerned, we see that Only 17% of resources report a documentation of some type, a finding that is in line with those reported in the introduction. The Size field was filled in by only 20% of entries and the Type of Licence is specified for only 18%.

### 3.2. Multidimensional Analysis

A multidimensional analysis is performed when the data are partitioned according to two or more dimensions. Combination of different metadata elements allows endless possibilities in analyzing the data. In this paper we concentrate in particular on exploring different dimensions starting from the "New" vs. "Old" Resources dichotomy, as this can be seen as one of the most innovative types of information brought by this Map.

#### 3.2.1. Resource Production Status and Resource Type

For instance, the data can be searched to find out whether the typology (according to the "Resource Type" element) of new resources differs or not from the one of existing resources. This is useful, since we can notice either if there are new trends in resource creation, or if the typology of newly created resources mirrors that of existing/used ones. Tables 8 and 9 illustrate the proportion of the various resource categories distinguishing between already existing and newly created ones.

| Resource Type | % |
|---|---|
| Corpus | 40.1 |
| Lexicon | 16.9 |
| Tagger/Parser | 16.2 |
| Annotation Tool | 7.9 |
| Ontology | 4.2 |
| Evaluation Tool | 1.8 |
| Terminology | 1.6 |
| Named Entity Recogniser | 1.6 |
| Evaluation Data | 1.6 |
| Language/Grammar Model | 1.3 |

Table 8: Types of existing resources

| Resource Type | % |
|---|---|
| Corpus | 53.2 |
| Lexicon | 10.8 |
| Annotation Tool | 8.0 |
| Ontology | 4.3 |
| Tagger/Parser | 4.1 |
| Representation-Annotation Formalism/Guidelines | 3.5 |
| Evaluation Data | 3.1 |
| Language/Grammar Model | 2.5 |
| Evaluation Tool | 2.0 |
| Named Entity Recogniser | 1.7 |
| Terminology | 1.6 |

Table 9: Types of newly created resources

#### 3.2.2. Resource Production Status and Resource Availability

It is interesting to see whether, and how, the availability status of newly created resources differs or not from already existing ones, and whether there is a variation according to their type. Figures 3 and 4 show that in comparison with already existing resources, new ones tend to be more directly available through the owner and less from Data Center, a finding that could reasonably be expected since many new resources are just completed or still in progress. Table 10 shows additional details. An unexpected finding, however, is that the proportion of freely available new resources is smaller than the one of freely available already existing resources. This might be due to the fact that new resources are not yet ready for distribution; in fact, if we further analyse this category by distinguishing between new resources still in progress and new resources already finished we can observe that the number of freely available completed new resources is almost as twice the number of resources that are still in progress (see Table 11).

The analysis can be further deepened by dividing between "Data-like" (corpora, lexicons, teminologies, ontologies, etc.) and "Tool-like" resources. Under this respect, it is interesting to observe that tools tend to be freely available more than data are (59% vs. 40%). This difference is mainly due to the availability of already existing tools: if we contrast new resources to already existing ones, we see that free availability of already existing tools increases to 69% vs. 45% for data, while the difference is not signifi-

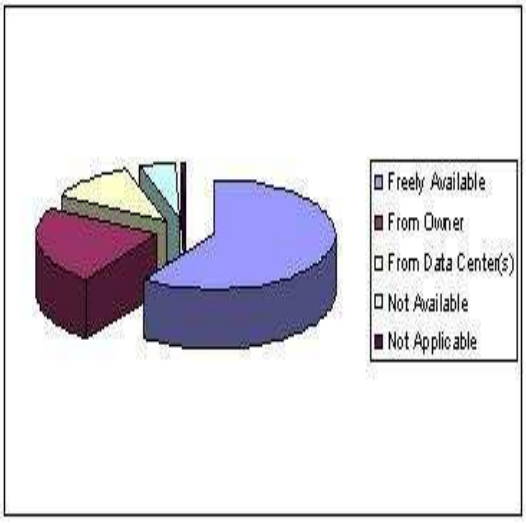


Figure 3: Availability status of already existing resources

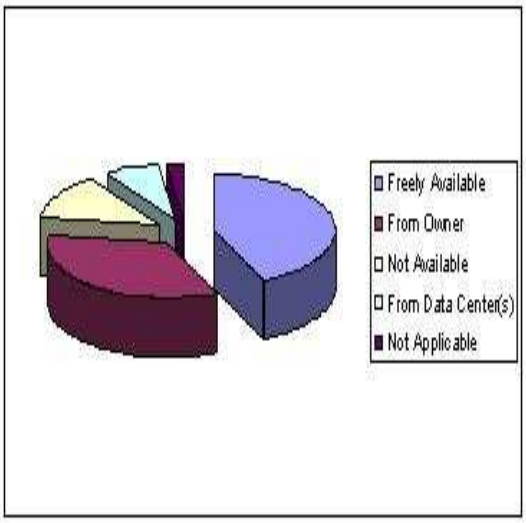


Figure 4: Availability status of new resources

cant for newly created resources (39% vs. 34%). See Table 12.

#### 3.2.3. Resource Production Status and Use of the Resource

An analysis similar to the one described in the previous section but more tailored to discovering new trends in use (either actual or intended) of new resources vs. older one can be done by combining the parameters "Resource Production Status" and "Use of the Resource". For both categories of resources the two most frequent uses are the same, i.e Information Extraction and Retrieval and Machine Translation. A difference starts to emerge from the third position, with the rising of applications such as Emotion and Dialogue for new resources. Strong application areas for already existing resources such as Word Sense Disambiguation or Question Answering do not even appear in the list of most frequent applications for new resources. See Table 13 for details.

| Availability | New | Existing |
|---|---|---|
| Freely available | 35% | 60% |
| From Owner | 28% | 23% |
| From Data Center | 6% | 11% |
| Not Available | 11% | 5% |
| Not Applicable | 2% | 1% |

Table 10: Availability status for newly created vs. already existing resources

| Availability | Finished | In progress |
|---|---|---|
| Freely available | 112 | 198 |
| From Owner | 90 | 155 |
| From Data Center | 20 | 32 |
| Not Available | 15 | 80 |
| Not Applicable | 3 | 14 |

Table 11: Availability status for finished vs. in progress newly created resources

#### 3.2.4. Other Multidimensional Analyses

As already told before, there are endless possibilities of analysis. Other interesting findings, such as how the various resources are distributed across languages and types, modality, etc. can also be derived. Coupling a Resource Type with the Language parameter can not only yield the amount of a given kind of resources for a certain language, but also the number of monolingual, bilingual and multilingual resources.

## 4. Using the Map

The LREC Map holds an unprecedented potential for possible applications and uses.

First, the Map is an instrument for enhancing availability of information about resources, either new or already existing ones, and it can be anticipated that one of its main uses will be as a cataloguing and searching facility. We expect that the Map will have a considerable impact: even before its official launch the idea was informally presented and circulated in the community, and other scientific communities asked for the possibility of replicating the information-getting procedure in conjunction with other conferences, scientific journals and events. When merged together, these different databases will form the most comprehensive repository of information about language resources to date, a "mother of all LR catalogs" that will enable simple and efficient access to information. On the community side, the Map represents an important preparatory step towards an open resource infrastructure, by allowing researchers and LR stakeholders to bottom-up provide information about language resources. After its launch at LREC, the interface will be migrated on a dedicated site and from there it will be accessible by anyone wishing to update a resource profile or adding a new one. By virtue of being a community-built and community-maintained repository of information, the Map is likely to reach existing resources in a more capillary way than usual catalogues.

The use of the Map as an information gathering tool is only

| | DATA | | TOOLS | |
|---|---|---|---|---|
| **Availability** | **New** | **Existing** | **New** | **Existing** |
| Freely available | 34.4 | 44.8 | 38.7 | 69.4 |
| From Data Center | 7.4 | 17.3 | 1.5 | 1.7 |
| From Owner | 26.6 | 23.3 | 31.4 | 18.9 |
| Not Available | 11.4 | 5.7 | 10.2 | 3.4 |
| Not Applicable | 2.0 | 0.5 | 1.5 | 1.0 |
| Other | 18.2 | 8.5 | 16.8 | 5.5 |

Table 12: Availability status for newly created vs. already existing resources

one of the many possible applications, and it is important to stress this point also to understand that there is no conflict between the Map and existing catalogues. Rather, the Map is an instrument conceived to complement and enhance them, also by acting as an instrument that will help permeating the community with entirely new conceptions about language resources and their documentation.

Apart from its use as a cataloguing and searching facility, another important use of the Map is as a measuring tool for monitoring various dimensions (metadata elements) of resources across places and times, thus helping highlighting evolutionary trends in language resource use and related human language technology development.

Finally, the potential "cultural" impact of the Map is probably most interesting. By cataloguing not only language resources in a narrow sense (i.e. language data), but also tools, standards, and annotation guidelines, it will help broadening the notion of "language resources" and thus attract to the field neighbouring disciplines that so far have been only marginally involved by the standard notion of LR.

By making most used/most adopted standards emerge, the Map will have an impact in reinforcing and facilitating the use of standards in the community. By allowing registration of resources together with submission of papers for a conference, it will pave the way to an entirely new tradition in the field of Language Resources and Technologies that ultimately may lead to the concept of publication and citation of language resources that may give academic credit along the lines of publications of papers.

Another application of the Map is related to metadata development and enhancement. By allowing user-defined metadata for describing resources and applications, the Map gives us interesting hints for assessing the usability and usefulness of a metadata set. It is interesting to see that while $80\%$ of entries used one of the descriptors provided, the remaining $20\%$ is represented by user-defined descriptors. An analysis of these two sets can also help in assessing the descriptive adequacy of a given metadata set. For instance, it appears that of the $20$ categories provided for classifying a resource according to its type, two were used only once (*speaker recogniser* and *language identifier*) and one (*metadata*) twice. On the other hand, some user-defined descriptors were independently used by different authors and recur more than two or three times, thus making good candidates for closed vocabularies in the future.

### 4.1. Resource Map as an Ontology

The LREC Map contains a big amount of hidden data, such as, for example, possible relations between resources, applications and between applications and resources. We can ask, thus, whether these data can be rendered as an ontology; and, as a consequence, what additional information can be extracted from the *ontologized* map. The above questions suggest a more important one:

> How should we design the ontology so that the hidden data in the LREC Map can be used to extract additional information ?

The basic idea, here, is to define as many ontological classes as the total number of distinct resource types extracted from the map. These classes are then grouped into more general super-classes. For example the *Corpus* resource type belongs to the *Resource-DataSet* super class[9]: according to this procedure, we can define the $is\_a$ relations type. In addition, we can specify that, for instance, a *Corpus* not simply $is\_a$ resource but it is a specific type of resource (a *Resource-DataSet*).

Once these classes have been defined, we can add individuals to them. Individuals are the resources being described that belong to a given resource type. Moreover, some relations can be extracted from the resource map data and used to formalize the classes. For example, the "used_for" and "used_by" relations can be extracted from the "resource use" metadata element.

## Acknowledgment

This work could not have been possible without the invaluable contribution of all the LREC authors, who patiently provided the input requested. The entire community is deeply indebted to them.

[9] Resource-DataSet, as well as Resource-Tool, Evaluation, ... have been defined for the Lrec Map tool described in section 2.2..

# Appendix

| Rank | Application of Already Existing Resources | Application of Newly Created Resources | Trend |
|---|---|---|---|
| 1 | Information Extraction, Information Retrieval | Information Extraction, Information Retrieval | ↔ |
| 2 | Machine Translation, SpeechToSpeech Translation | Machine Translation, SpeechToSpeech Translation | ↔ |
| 3 | Knowledge Discovery/Representation | Emotion Recognition/Generation | ↑ |
| 4 | Language Modelling | Knowledge Discovery/Representation | ↓ |
| 5 | Acquisition | Acquisition | ↔ |
| 6 | Word Sense Disambiguation | Dialogue | ↑ |
| 7 | Named Entity Recognition | Discourse | ↑ |
| 8 | Discourse | Language Modelling | ↓ |
| 9 | Question Answering | Language Identification | ↑ |
| 10 | Document Classification, Text categorisation | Named Entity Recognition | ↓ |

Table 13: A comparison of uses of existing vs. new resources