LRE Map, a Song of Resources and Evaluation

Riccardo Del Gratta, Sara Goggi, Gabriella Pardelli, Nicoletta Calzolari
Istituto di Linguistica Computazionale “A. Zampolli”, CNR
Via Moruzzi 1, 56124 Pisa, Italy
{firstname.lastname}@ilc.cnr.it

Abstract

After 8 years we revisit the LRE Map of Language Resources, introduced at LREC 2010, to try to get a picture of the field and its evolution as reflected by the creation and use of Language Resources. The purpose of the Map was in fact “to shed light on the vast amount of resources that represent the background of the research presented at LREC”. It also aimed at a “change of culture in the field, actively engaging each researcher in the documentation task about resources”. The data analysed here have been provided by the authors of several conferences during the phase of submission of papers, and contain information about ca. 7500 resources. We analysed the LRE Map data from many different viewpoints and the paper reports on the global picture, on different trends emerging from the diachronic perspective and finally on some comparisons between the 2 major conferences present in the Map: LREC and COLING.

Keywords: LR Infratructure, Metadata, LR Documentation

1. The LRE Map and its data

1.1. The LRE Map: why?

Science is ever more driven by data and our field is not different. Natural Language Processing (NLP) is certainly a data-intensive field. The LRE Map of Language Resources[1] (data and tools) was an innovative instrument introduced in the LREC2010 conference (Calzolari et al., 2010) with the aim of monitoring and representing the wealth of data and technologies developed and used in the field.

We called it “Map” because we aimed at representing the relevant features of a large territory, also for the parts not represented in the official catalogues of major players of the field (ELRA, LDC, NICT, ACL, OLAC, LT World, etc.). But we had other purposes too: we wanted to draw attention to the importance of the Language Resources (LRs) that are behind many of our papers; we wanted to start giving the deserved recognition, as suggested by the FLaReNet Thematic Network recommendations (Soria et al., 2014), to the developers of Language Resources (LRs); and finally wanted to map the “use” of LRs, to understand the purposes of the developed LRs and how their intended usage changes over time.

The collaborative creation of the Map was undoubtedly critical: we conceived the Map as a means to influence a “change of culture” in our community, whereby everyone is asked to make a minimal effort to document the LRs that are used or created. By spreading the LR documentation effort across many people instead of leaving it only in the hands of the distribution centres, we also encourage awareness of the importance of metadata and proper documentation. Documenting a LR is the first step towards identifiability, which in its turn is the first step towards reproducibility.

We kept the requested information at a simple level, knowing that we had to compromise between richness of metadata and willingness of authors to fill them in.

With all these purposes in mind we thought we could exploit the great opportunity offered by LREC and the involvement of so many authors from so many countries, from different modalities and working in so many areas of NLP.

Afterwards the Map was used also in the framework of other major Conferences, in particular by COLING, and this provides another opportunity for useful comparisons. It was an innovative action of infrastructural nature, in the awareness that research is affected also by such activities.

1.2. The LRE Map: the current data

We provide here a general overview of the data collected so far. The total number of LRs described in the Map is 7453 (instances), collected from 17 different conferences[2] (some with workshops). The major conferences for which we regularly have data are LREC and COLING.

The set of metadata to be used for describing each LR is the following: *Resource Type, Name, Production Status, Use, Language, Modality, Availability, Size, License, Documentation, URL, Description*. This set is very basic and is compatible with the metadata of other major LR catalogues, such as ELRA, CLARIN, META-SHARE.

The Map contains information also on its contributors and their organisations/institutions, countries, etc. We do not analyse these data here (also because not normalised), but they could be the source of other types of interesting analyses in the future.

Let us see first how LRs are distributed along some of these dimensions (considering only the most frequent values). As concerns LR Types, they are 469 in total, this variety being due to the possibility for authors of inserting a free description of the LR type in addition to those provided by us.

[1] The LRE Map, currently being updated, is at `http://lremap.elra.info` or `http://www.resourcebook.eu`

[2] This is the list of all conferences: LREC 10-12-14-16, COLING 10-12-14-16, INTERSPEECH 11-13, IJCNLP 11, RANLP 11, ACLHLT 11, LTC 11, O-COCOSDA 11, NAACL 13, BioTxtM 14.

Current figures are reported in the following tables. Table 1 lists the most frequent LR Types; Table 2 the most frequent modalities where *Written* is by large the most frequent, which is obvious given the type of conferences (even if at LREC we aim to have all the modalities represented).

LR Type	**Percentage**
corpus	46.48
lexicon	11.2
tagger/parser	6.63
annotation tool	4.12
evaluation data	3.65
ontology	2.79
corpus tool	1.73
.....	....

Table 1: Percentage of the 7 most frequent Types

LR Modality	**Percentage**
written	72.32
speech	6.45
multimodal/multimedia	4.39
not applicable	4.05
speech/written	3.01
modality independent	1.01
sign language	0.95

Table 2: Percentage of Modalities

Table 3 displays the LR Availability, and we remark that less than 5% of the LRs are not available at all; Table 4 shows figures for the LR Production Status and hints that there is a similar distribution between already existing LRs and newly created ones.

LR Availability	**Percentage**
freely available	50.23
from owner	20.73
from data center(s)	8.65
not available	4.56

Table 3: Percentage of Availability

LR Status	**Percentage**
existing-used	41.07
newly created-in progress	24.04
newly created-finished	20.81
existing-updated	8.16

Table 4: Percentage of Resource Production Status

Table 5 provides the major Usages, which is an indicator of where the efforts of our community are mostly concentrated in these years.

LR Usage	**Percentage**
information extraction, information retrieval	10.08
machine translation, speechtospeech translation	8.39
parsing and tagging	4.96
language modelling	4.49
document classification, text categorisation	3.17
evaluation/validation	2.94
knowledge discovery/representation	2.9
corpus creation/annotation	2.87
acquisition	2.78
speech recognition/understanding	2.74
discourse	2.62
named entity recognition	2.56
.....	....

Table 5: Percentage of the 12 most frequent Usages

1.3. Metadata values and their normalisation

This section refers to LRE Map metadada normalisation as part of the curation process on the Map data (Del Gratta et al., 2014).

In the tables above we provided only the most frequent values, but there are many “long tails” with small numbers or single occurrences (hapax), due to the possibility for authors to introduce a personal value under “other”. After the first LREC we started to normalise some of the most frequent occurrences of personal values and introduced some of them among the “suggested values”.

Metadata normalisation is an important (and never ending) process to reduce redundancies and ensure better accuracy. Simple examples of the normalisation process of authors’ metadata values are: the pre-processing of values to eliminate the possible sources of duplicate information by standardizing the different spelling of proper names, acronyms and other terms used by the authors; insertion of the ISO code for languages; solving acronyms.

The strategy we have followed to address the above issues is to carry on a normalisation process of the values provided by the authors and to link the original values to the normalised ones. For us it is important to keep also the original information provided by the authors because of the bottom-up nature of the LRE Map.

We must also underline that in the Map there are, obviously, many cases of reference to the same Language Resource. Differently from other catalogues, in the Map it is important to know how much a LR is used and cited; for this reason we allow multiple occurrences/descriptions of the same LR. Normalisation of LR Names is therefore important to allow for grouping the different mentions of the same LR under the same name. This way we know which are the most “used/cited” LRs.

2. Some trends

2.1. General evolution

After 8 years of use, it is interesting to see whether there has been an evolution in the creation and use of Language Resources: we introduce here some tables that show some variability in the use of different LRs along the years.

2.2. LR Types

Table A [3] gives the distribution per year of the most frequent LR Types, showing with arrows the trend over the previous year: corpora and evaluation data are constantly increasing while tagger/parser, annotation tool and ontology are significantly decreasing.
In addition Figure I-a plots the data distribution of Types with very different distribution over the years.[4]

2.3. LR Availability

Table B shows that free-availability of LRs has greatly increased during these years: this is an important remark that reflects how the field is evolving with respect to an increasing awareness of the value of data sharing and openness.

2.4. LR Usage

Table C shows some interesting differences in the usage of LRs along the years. Information Extraction/Information Retrieval and Machine Translation, that were constantly the first usages until 2014, go down remarkably in 2016. While very interestingly many more LRs are used for evaluation purposes. This is a clear sign of the increasing importance attributed to evaluation in our field.

2.5. LR Status

Table D shows that there is a tendency towards creating new LRs while the re-use of existing ones is slowly decreasing.

3. Multi-dimensional analysis

We can take into account combinations of different metadata to look at various correlations, with a multi-dimensional analysis. This analysis provides some of the most interesting results.

3.1. Correlation between Type and Status

We see here the correlation between LR Type and LR Status, i.e. which Types of LRs are mostly newly created or existing and used.
There are some LR Types that behave really different from the average trend: Evaluation data are much more Newly built, while exactly the contrary is for Tagger/Parser and for Named Entity Recognizer that are mostly Existing and re-used (see Table 6). There is apparently no longer a great need of developing new taggers/parses or NE recognisers. While, coherently with the observation just made above about the increasing importance of evaluation data, there is the need of more and new Evaluation data.

LR Type vs. Status	% Existing	% New
Total LRs	49.26	44.8
Evaluation Data	38.99	61.01
Tagger/Parser	78.41	21.58
Named Entity Recognizer	63.63	36.37

Table 6: Correlation between Type and Status

3.2. Correlation between Availability and Modality and Status

From Table E it appears very clearly that Written Language Resources are more freely available than Spoken and Multimodal LRs, while these are more frequently available from the owner with respect to the others. Spoken LRs are comparatively more distributed through data centers. And Spoken and Multimodal (and also Sign language) LRs are the most not-available. From these data the Written community seems to be more willing to go in the direction of openness. But we must observe that Spoken and Multimodal LRs are usually much more expensive to create.
Among the freely-available LRs the existing-used/updated ones are more numerous than the new ones. The new-in-progress are comparatively (and obviously) more available from owner and also the most frequent in the not-available (hopefully just not yet). Data centers clearly distribute mostly the existing-used LRs.

4. Comparison between LREC and COLING: some interesting differences

It is interesting to see which are the major features that characterise different conferences with respect to Language Resources. The Map helps us in this, in particular to look at the differences between LREC and COLING for which we collected information for the last 4 editions of both conferences (2010, 2012, 2014, 2016).
There is a clear difference with respect to LR Modality, as reflected in Figure 1: LREC has more participation of the Speech and Multimodal communities and therefore more Speech and Multimodal LRs. A typical feature of LREC is that the Sign Language community is well represented (and consequently the corresponding type of LRs), while these are completely absent from COLING.
Figure 3 shows the comparison of LR Usages: the three most frequent Usages are exactly the same, in the same order, with similar frequency. But others have different frequencies in the two conferences.
Interestingly different is the Status of LRs: at LREC there are many more new Language Resource while at COLING more existing ones (Figure 2). This is expected given the nature of the conferences and the specific focus of LREC on LRs.

[3] Additional tables and figures can be found in the Appendix.

[4] Other plots on the LR Type data distribution are available at http://www.resourcebook.eu/trends/lr_years.html.

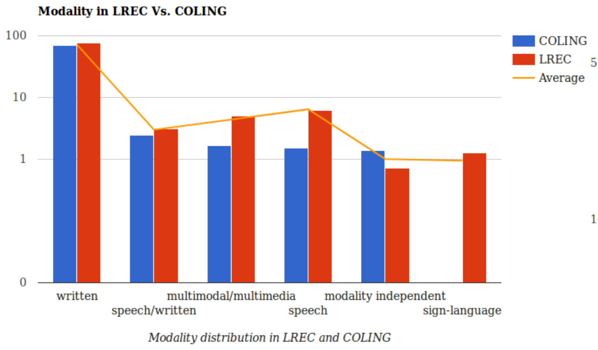

Figure 1: LR Modality LREC vs.COLING

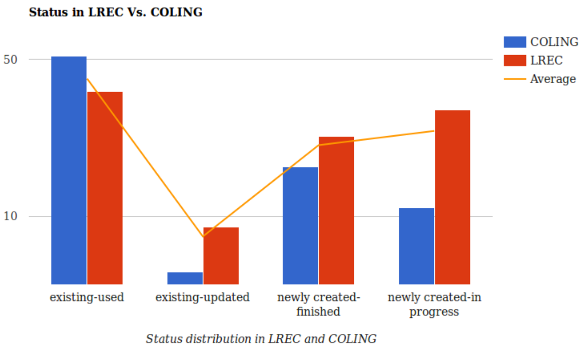

Figure 2: LR Status LREC vs.COLING

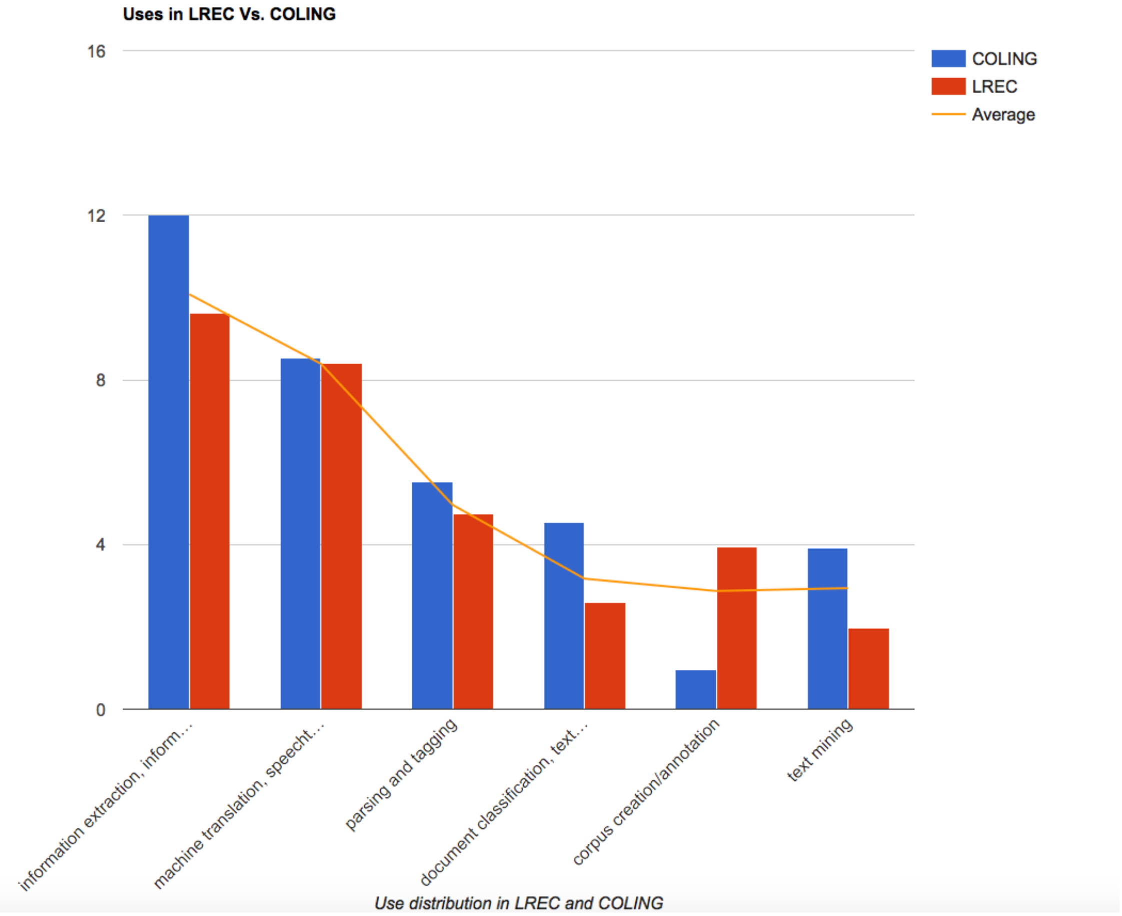

Figure 3: LR Use LREC vs.COLING

5. Conclusions

With initiatives such as the LRE Map and "Share your Language Resources" (introduced in 2014) we want to encourage in the field of Language Resources and Language Technology what is already in use in more mature disciplines, i.e. proper documentation and reproducibility as a normal practice. We think that research is strongly affected also by such infrastructural (meta-research) activities and therefore we continue to promote - also through such initiatives - a greater visibility of LRs, the sharing of LRs in an easier way and the reproducibility of research results.

Here is the vision: it must become common practice also in our field that when you submit a paper either to a conference or a journal you are offered the opportunity to document and upload the LRs related to your research. This is even more important in a data-intensive discipline like NLP. The small cost that each of us will pay to document, share, etc. should be paid back from benefiting of others' efforts.

6. Bibliographical References

Arranz, V., Choukri, K., Mapelli, V., and Mazo, H. (2014). ELRA's Consolidated Services for the HLT Community. In Nicoletta Calzolari (Conference Chair), et al., editors, *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 1511–1516, Reykjavik, Iceland. European Language Resources Association (ELRA).

Calzolari, N., Soria, C., Del Gratta , R., Goggi, S., Quochi, V., Russo, I., Choukri, K., Mariani, J., and Piperidis, S. (2010). The LREC Map of Language Resources and Technologies. In Nicoletta Calzolari (Conference Chair), et al., editors, *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, pages 949–956, Valletta, Malta. European Language Resources Association (ELRA).

Calzolari, N., Del Gratta, R., Francopoulo, G., Mariani, J., Rubino, F., Russo, I., and Soria, C. (2012). The LRE Map. Harmonising Community Descriptions of Resources. In Nicoletta Calzolari (Conference Chair), et al., editors, *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 1084–1089, Istanbul, Turkey. European Language Resources Association (ELRA).

Cieri, C., Choukri, K., Calzolari, N., Langendoen, D. T., Leveling, J., Palmer, M., Ide, N., and Pustejovsky, J. (2010). A Road Map for Interoperable Language Resource Metadata. In Nicoletta Calzolari (Conference Chair), et al., editors, *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, pages 2506–2509, Valletta, Malta. European Language Resources Association (ELRA).

Del Gratta, R., Pardelli, G., and Goggi, S. (2014). The LRE Map disclosed. In Nicoletta Calzolari (Conference Chair), et al., editors, *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 3534–3541, Reykjavik, Iceland. European Language Resources Association (ELRA).

Del Gratta, R., Frontini, F., Monachini, M., Pardelli, G., Russo, I., Bartolini, R., Khan, F., Soria, C., and Calzolari, N. (2016). LREC as a Graph: People and Resources in a Network. In Nicoletta Calzolari (Conference Chair), et al., editors, *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, pages 2529–2532, Portorož, Slovenia. European Language Resources Association (ELRA).

Francopoulo, G., Mariani, J., and Paroubek, P. (2016). Predictive Modeling: Guessing the NLP Terms of Tomorrow. In Nicoletta Calzolari (Conference Chair), et al., editors, *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, pages 336–343, Portorož, Slovenia. European Language Resources Association (ELRA).

Mapelli, V., Arranz, V., Carré, M., Mazo, H., Mostefa, D., and Choukri, K. (2012). ELRA in the heart of a cooperative HLT world. In Nicoletta Calzolari (Conference Chair), et al., editors, *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, pages 55–59, Istanbul, Turkey. European Language Resources Association (ELRA).

Mariani, J., Paroubek, P., Francopoulo, G., and Hamon, O. (2014). Rediscovering 15 Years of Discoveries in Language Resources and Evaluation: The LREC Anthology Analysis. In Nicoletta Calzolari (Conference Chair), et al., editors, *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 4632–4669, Reykjavik, Iceland. European Language Resources Association (ELRA).

Mariani, J., Paroubek, P., Francopoulo, G., and Hamon, O. (2016). Rediscovering 15 + 2 years of Discoveries in Language Resources and Evaluation. *Language Resources and Evaluation*, 50(2):165–220.

Soria, C., Calzolari, N., Monachini, M., Quochi, V., Bel, N., Choukri, K., Mariani, J., Odijk, J., and Piperidis, S. (2014). The Language Resources Strategic Agenda: the FLaReNet synthesis of community recommendations. *Language Resources and Evaluation*, 48(4):753–775.

Wikipedia. (2017). LRE Map — Wikipedia, The Free Encyclopedia. [Online; accessed 2-October-2017].

7. Language Resource References

Nicoletta Calzolari and Claudia Soria and Riccardo Del Gratta and Sara Goggi and Valeria Quochi and Irene Russo and Khalid Choukri and Joseph Mariani and Stelios Piperidis. (2010). *The LRE Map*. ELRA/ILC.

Appendix: Tables and Figures

LR Type	%2010	%2012	%2014	%2016
corpus	41.37	46.16 (↑)	48.9 (↑)	49.12 (↔)
lexicon	11.5	11.86 (↔)	11.8 (↔)	10.89 (↓)
tagger/parser	9.72	5.78 (↓)	3.74 (↓)	4.47 (↑)
annotation tool	5.29	4.33 (↓)	3.90 (↓)	2.71 (↓)
evaluation data	3.36	3.80 (↔)	2.72 (↓)	4.41 (↑)
ontology	3.36	3.12 (↔)	3.14 (↔)	1.70 (↓)
...	...	...	...	...

Table A: Percentage of the first 6 LR Types with arrows showing the trend wrt previous year

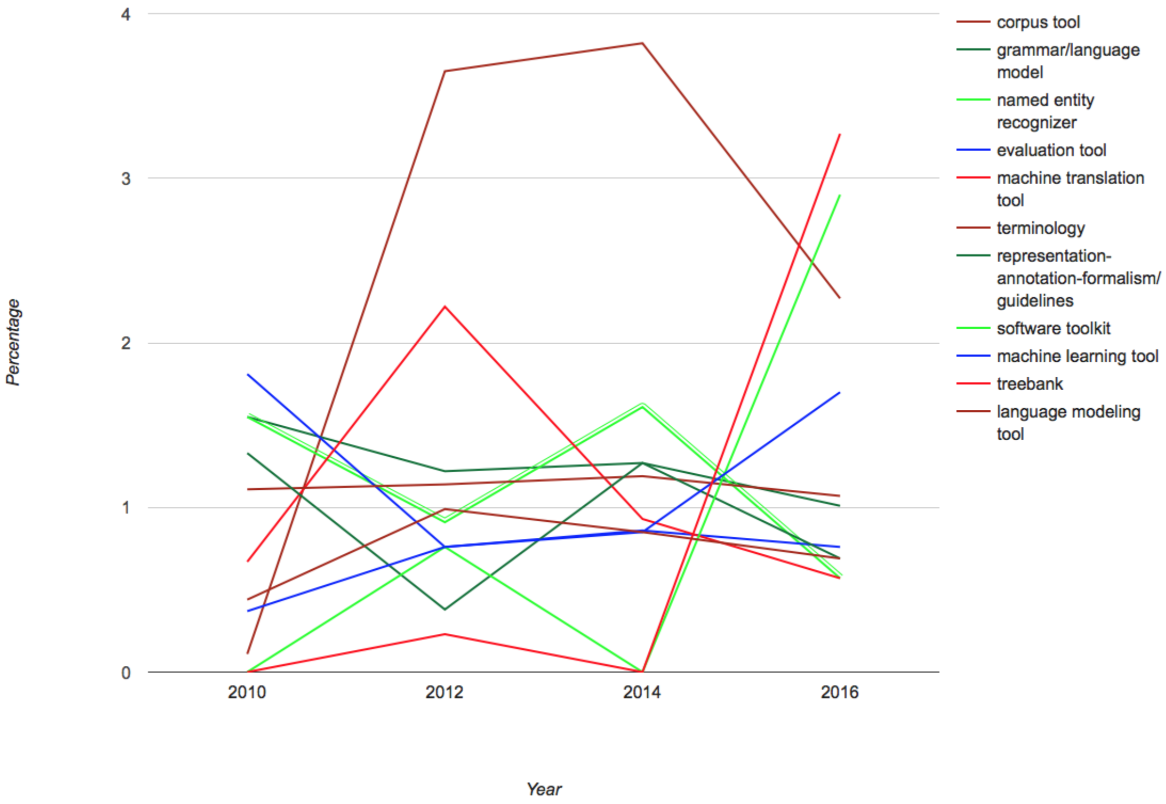

Figure I-a: Plot of volatile data

LR Availability	%2010	%2012	%2014	%2016
freely available	44.88	52.7 (↑)	55.4 (↑)	57.3 (↑)
from owner	22.22	19.77 (↔)	20.88 (↑)	19.71 (↓)
from data center(s)	9.09	8.21 (↓)	6.03 (↓)	7.3 (↑)
not available	5.73	3.95 (↓)	3.99 (↔)	3.21 (↔)

Table B: Percentage of Availability with arrows showing the trend wrt previous year

LR Usage	%2010	%2012	%2014	%2016
information extraction, information retrieval	11.61	12.09 (↔)	10.27 (↓)	6.3 (↓)
evaluation/validation	1.37	1.6 (↔)	1.53 (↔)	8.63 (↑)
machine translation, speechtospeech translation	9.17	10.49 (↑)	8.57 (↓)	5.35 (↓)
knowledge discovery/ representation	4.25	2.05 (↓)	2.04 (↔)	2.39 (↔)
...	...	...	...	...

Table C: Percentage of main Usages with arrows showing the trend wrt previous year

LR Status	%2010	%2012	%2014	%2016
existing-used	40.06	39.85 (↔)	27.42(↓)	34.89 (↑)
newly created-in progress	24.33	23.19 (↓)	31.92 (↑)	22.61 (↓)
newly created-finished	13.12	21.98 (↑)	26.99 (↑)	29.79 (↑)
existing-updated	8.17	8.52 (↔)	9.68 (↑)	6.86 (↓)

Table D: Percentage of Status with arrows showing the trend wrt previous year

Availability	Status (%)			
	Existing-used	Existing-updated	Newly created-finished	Newly created-in progress
Freely Available	61.06	64.96	57.74	48.76
From Owner	20.56	23.23	29.58	33.29
From Data Center(s)	15, 93	7, 68	7, 4	5, 71
Not Available	2, 45	4, 13	5, 28	12, 24

Table E: Availability Vs. Status

Availability	Modality (%)					
	Written	Speech/ Written	Speech	Multimodal/ Multimedia	Sign Language	Modality Independent
Freely Available	60.18	48.07	30.92	41.04	40.00	77.36
From Owner	24.57	29.83	36.16	36.94	36.36	20.75
From Data Center(s)	10.34	16.57	21.7	9.7	5.45	1.89
Not Available	4.91	5.52	11.22	12.31	18.18	

Table F: Availability Vs. Modality