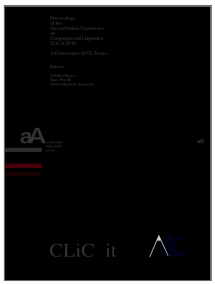

Cristina Bosco, Sara Tonelli and Fabio Massimo Zanzotto (dir.)

**Proceedings of the Second Italian Conference on Computational Linguistics CLiC-it 2015**
**3-4 December 2015, Trento**

Accademia University Press

# Visualising Italian Language Resources: a Snapshot

**Riccardo Del Gratta, Francesca Frontini, Monica Monachini, Gabriella Pardelli, Irene Russo, Roberto Bartolini, Sara Goggi, Fahad Khan, Valeria Quochi, Claudia Soria and Nicoletta Calzolari**

DOI: 10.4000/books.aaccademia.1442
Publisher: Accademia University Press
Place of publication: Torino
Year of publication: 2015
Published on OpenEdition Books: November 11, 2016
Series: Collana dell'Associazione Italiana di Linguistica Computazionale
Electronic EAN: 9788899200008

# Visualising Italian Language Resources: a Snapshot

**Riccardo Del Gratta, Francesca Frontini, Monica Monachini, Gabriella Pardelli, Irene Russo, Roberto Bartolini, Sara Goggi, Fahad Khan, Valeria Quochi, Claudia Soria, Nicoletta Calzolari**
Istituto di Linguistica Computazionale "A . Zampolli"
CNR Pisa, Italy
name.surname@ilc.cnr.it

## Abstract

**English.** This paper aims to provide a first snapshot of Italian Language Resources (LRs) and their uses by the community, as documented by the papers presented at two different conferences, LREC2014 and CLiC-it 2014. The data of the former were drawn from the LOD version of the LRE Map, while those of the latter come from manually analyzing the proceedings. The results are presented in the form of visual graphs and confirm the initial hypothesis that Italian LRs require concrete actions to enhance their visibility.

**Italiano.** *Questo articolo ha l'obiettivo di fornire una fotografia del contesto delle Risorse Linguistiche italiane e dei loro usi da parte della comunità scientifica; i dati usati sono tratti dagli articoli presentati a due diverse conferenze del settore, LREC2014 e CLiC-it 2014. I primi sono derivati dalla LRE Map in versione LOD, mentre i secondi sono stati ottenuti da un'analisi manuale degli atti della conferenza. I risultati sono presentati e analizzati sotto forma di grafi e confermano l'ipotesi che le risorse linguistiche italiane richiedano azioni mirate ad aumentare la loro visibilità.*

## 1 Introduction

The availability of Language Resources (LRs) - such as corpora, computational lexicons, parsers, etc. - is crucial to most NLP technologies (Machine Translation, Crosslingual Information Retrieval, Multilingual Information Extraction, Automatic Document Indexing, Question Answering, Natural Language Interfaces, etc.). Recent initiatives have monitored the availability of language resources for different languages, and highlighted a digital divide between English and other languages (Soria et al., 2012). While the economic potential of English ensures that English LRs are developed and maintained not only in the academic sector but also by commercial players, the involvement of research communities for languages such as Italian is much more crucial to ensure that the necessary instruments (both data and tools) are made available for natural language processing purposes.

At the same time, the production of quality LRs is just a first step; LRs must also be documented and made available to the community in such a way that they are easy to find and to use. This entails the description of every LR with a set of metadata that clarify its typology, its language, its size and licensing scheme, and the means of accessing it. Useful information in this sense can be found in the catalogues of language resources associations, such as ELRA, LDC, NICT Universal Catalogue, ACL Data and Code Repository, OLAC, LT World. These catalogues adopt a top-down approach to documenting resources and typically list resources that have reached a high level of maturity - in term of validation, documentation, clearing of IPR issues, etc. As an alternative to this approach, recent projects have been carried out within the LR community to create open, bottom-up repositories where LRs - even those under development - can be duly documented and searched. Such initiatives are for instance the META-SHARE platform (Gavrilidou et al., 2012), the CLARIN VLO (Broeder et al., 2010) and the LRE Map (Calzolari et al., 2012; Del Gratta et al., 2014b; Del Gratta et al., 2014a), with their sets of metadata. In particular the LRE Map was launched as an initiative at LREC2010 in order to crowdsource reliable and accurate documentation for the largest possible set of resources. Au-

thors submitting to that conference were asked to document the resources they used in their paper, both the resources they created and the ones created by others. This initiative has continued and been extended to other conferences[1], and is now a unique source of information on existing language resources and their use in current research. The work in this paper can be set against the background of the major projects in which CNR ILC is currently involved and the aim of setting up a documentation center for language and textual resources within the framework of the CLARIN and DARIAH research infrastructures. As a CLARIN and DARIAH node, CNR ILC has the task of collecting and harmonizing metadata description of LRs at a national level, making Italian resources more visible to national and international research groups, both to the NLP and to the digital humanities communities. To this purpose, our team has inspected the panorama of LR descriptions available in the aforementioned catalogues, and in particular the LRE Map which allows us to monitor how communities build around LR use. Our hypothesis is that many of the resources that the Italian community uses and produces are not as well documented as they should be. As a consequence, many researchers may not be aware of the existence of resources that could be of use for them, and limit themselves to those they know best. In order to verify this, we carried out a cross-analysis of Italian LRs and their uses by Italian researchers, exploiting the data found in the LRE Map from the LREC2014 dataset, which is currently available in LOD format (Del Gratta et al., 2014a). Such data is compared with similar evidence gathered from the proceedings of the CLiC-it 2014 conference, which are available online. CLiC-it 2014 did not adhere to the LRE Map initiative, but comparable information has been collected by manually inspecting the papers. In what follows we will provide a brief description of the set of metadata that we used to monitor the situation with respect to Italian LRs and their use; then some results will be analyzed and discussed by means of graph-like visualizations; finally some conclusions are drawn and perspectives for future work outlined.

[1]Such as COLING, EMNLP, ACL-HLT, RANLP, Interspeech, Oriental-Cocosda, IJCNLP, LTC, NA-ACL

## 2 Metadata description

The set of metadata used for documenting language resources can vary from repository to repository. Some harmonization initiatives are currently being carried out in order to make diverse datasets interoperable, e.g. (McCrae et al., 2015). Nevertheless a common core has been broadly agreed upon by all; this includes type of resource (corpus, lexicon, tool), modality, language(s), use, availability. To this core set of metadata, the LRE Map adds other metadata that are linked not to the resource itself, but to its use in the paper that is being submitted: thus information about the conference, the paper, the authors and their affiliations is available for each entry in the LRE Map. This also means that any given resource can have more than just one entry in the LRE Map, one for each paper that has used it. Sometimes the resource is marked as new, and in that case we can assume that the authors of the paper are also the producers of this new resource; in most cases the resource is a well known one. So for instance some of the most used resources according to the LRE Map are Princeton WordNet and the British National Corpus. For the purposes of this paper we only took into consideration the following metadata for each entry in the LREC2014 LRE Map: resource name, language, authors and affiliations. We extracted all used LRs with Italian as one of the languages and authors with an Italian affiliation. We then analysed the proceedings of CLiC-it 2014 and manually extracted the same type of information for each paper [2]. We thus obtained two datasets:

Table 1: LRs use - the Italian panorama.

| | Authors | LRs | Institutions | Papers |
|---|---|---|---|---|
| LREC '14 | 91 | 25 | 41 | 24 |
| CLiC-it '14 | 107 | 54 | 28 | 42 |
| Total '14 | 166 | 74 | 57 | 66 |

[2]One of the most interesting features of the LRE Map is the fact that it provides a user's perspective on language resources. So for instance Princeton WordNet may be defined by some as a lexicon and by others as an ontology; moreover the declared use may vary from paper to paper. In the case of the CLiC-it dataset the data was collected by just one person, and thus this precious information is not available. For this reason this data cannot be inserted into the LRE Map and has to be considered as a simulation.

## 3 People and Resources: visualising networks

Data visualisation is a method that enables the exploration, filtering and searching of data, skipping the interaction with databases. Data can be mainly visualised for presentation or exploration but in well designed projects there is a continuum between these two modalities (Cairo, 2013).
In this paper we propose two visualisation modalities to discover the interrelations between authors from different institutions and the convergence of authors on the usage of the same resource. In comparing these two conferences the aim was to portray the Italian NLP community highlighting collaborations between people through resources used.
The implementation of the visualisation is based on a well known tool, D3.js, a JavaScript library designed to display digital data in a dynamic graphical form. The two visualisations are:

- a force-directed graph (see a detail in Figure 1)[3] where each author is a node; the links between author-nodes stand for co-authorship in a paper. Different institutions are assigned different colours; in this way people belonging to the same institution are visually identifiable and collaborations among institutions are clear because of the links connecting coauthors of different colours: for example Cristina Bosco from the University of Turin is connected to co-authors from the same institution (purple dots) but also to Maria Simi from the University of Pisa and Simonetta Montemagni from ILC CNR (orange and brown dot, respectively).

- a force-directed graph where each author is a node connected to other persons only through the resources they use, depicted as boxes. Here too, the colour of the person depends on the institution. People are connected to the same resource (1) when they co-authored a paper that uses it, (2) because they use the same resource in independent research works. In the first case, co-author groups are still somewhat identifiable, as they create an island effect (as shown in Figure 2). In the other case heterogeneous people get connected because they use the same resources. As a result, networks of researchers are gathered around LR uses (see Figure 3).

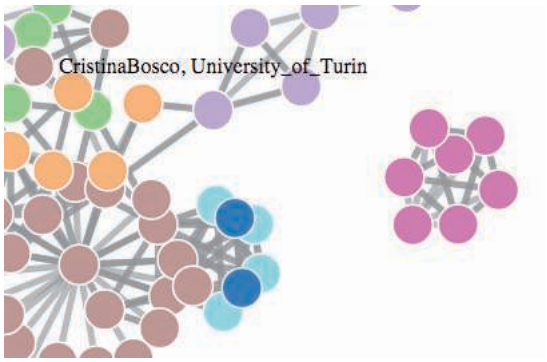


Figure 1: Cross institution co-author networks.

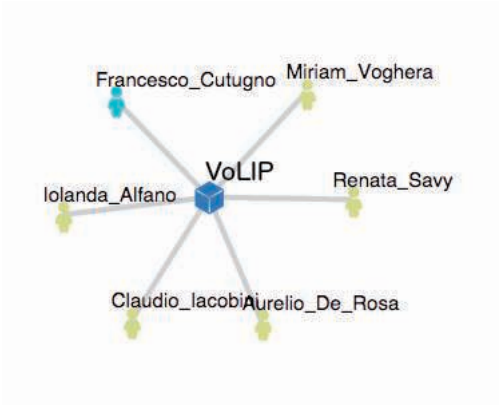


Figure 2: Same resource used by co-authored paper.

Graph-based visualisations pave the way for a social network analysis of the data that we plan as future work. For the moment, thanks to these two graphical devices, some interesting phenomena are now visually evident; we concentrate in particular on how research collaborations gather around LRs. The first phenomenon is that at the LREC2014 there are more international collaborations between Italian and foreign groups. The first edition of CLiC-it instead presents less involvement of foreign co-authors and more collaborations between different Italian institutions. This is clearly due to the fact that CLiC-it is a national conference, while LREC an international one. The second fact is that at LREC2014 we find a smaller number of Italian LRs, as typically papers use the best known ones. CLiC-it instead presents us with a broader panorama: in addition to the best known resources we find a plethora of minor resources -

[3]The interactive visualisations are available online at http://www.clarin-it.it/jvis

in particular corpora - that are not mentioned in the LREC2014 dataset and are mostly used in a single paper. In many cases the user of the resource is also its creator: these resources need documentations to foster future collaborations. Graph-based

## 4 Conclusions and future works

In this work we use visualisations to show how the Italian NLP community uses LRs in the works presented at two recent conferences of the sector (LREC2014 and CLiC-it 2014). We highlight how collaborations cluster around the use of major resources, and how networks are created by users of the same resource. From the comparison of the two datasets we can infer that the Italian panorama of language resources is rich and varied. We also confirm the prior hypothesis that Italian LRs are rather under-documented and that some positive action is needed in the direction of enhancing their visibility. As a consequence the creation of an observatory of Italian language resources, which is meant to be the nucleus of a newly established CLARIN-IT center, is more than justified. Such an observatory will actively promote the Italian LR community (both creators and users), help in improving the documentation of LRs thus making them more widely known to others and finally ensure their visibility in an international context by using all current standard metadata framework and platforms. This latter point shall involve also an active contribution to the de-fragmentation of the current situation in metadata and description practices, as well as the porting of LR descriptions to emerging channels and formats (LINGhub[4], RDF-LOD).

## Acknowledgments

The research carried out in this paper was partly funded by SM@RTINFRA (MIUR Progetto premiale) and PARTHENOS (H2020 INFRADEV-4).

## References

Daan Broeder, Marc Kemps-Snijders, Dieter Van Uytvanck, Menzo Windhouwer, Peter Withers, Peter Wittenburg, and Claus Zinn. 2010. A data category registry-and component-based metadata framework. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, pages 43–47. European Language Resources Association (ELRA).

Alberto Cairo. 2013. *L'arte funzionale: Infografica e visualizzazione delle informazion*. Pearson Italia Spa.

Nicoletta Calzolari, Riccardo Del Gratta, Gil Francopoulo, Joseph Mariani, Francesco Rubino, Irene Russo, and Claudia Soria. 2012. The LRE Map. Harmonising Community Descriptions of Resources. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 1084–1089. European Language Resources Association (ELRA).

Riccardo Del Gratta, Francesca Frontini, A Fahad Khan, Joseph Mariani, and Claudia Soria. 2014a. The LRE Map for under-resourced languages. In *Workshop Collaboration and Computing for Under-Resourced Languages in the Linked Open Data Era, Satellite Workshop of LREC'14*.

Riccardo Del Gratta, F Khan, Sara Goggi, and G Pardelli. 2014b. LRE Map disclosed. In *Proceedings of the ninth International Conference on Language Resources and Evaluation (LREC'14)*. European Language Resources Association (ELRA).

Maria Gavrilidou, Penny Labropoulou, Elina Desipri, Stelios Piperidis, Harris Papageorgiou, Monica Monachini, Francesca Frontini, Thierry Declerck, Gil Francopoulo, Victoria Arranz, et al. 2012. The META-SHARE Metadata Schema for the Description of Language Resources. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 1090–1097. European Language Resources Association (ELRA).

John McCrae, Penny Labropoulou, Jorge Gracia, Marta Villegas, Victor Rodriguez-Doncel, and Philipp Cimiano. 2015. One ontology to bind them all: The META-SHARE OWL ontology for the interoperability of linguistic datasets on the Web. In *Proceedings of the 4th Workshop on the Multilingual Semantic Web*.

Claudia Soria, Nria Bel, Khalid Choukri, Joseph Mariani, Monica Monachini, Jan Odijk, Stelios Piperidis, Valeria Quochi, Nicoletta Calzolari, and others. 2012. The FLaReNet Strategic Language Resource Agenda. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 1379–1386. European Language Resources Association (ELRA).

[4] http://linghub.lider-project.eu/

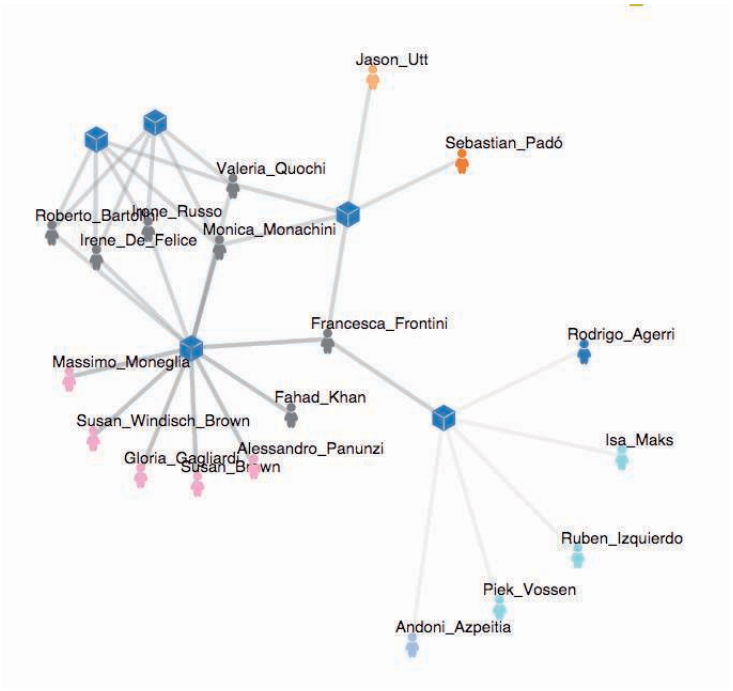


Figure 3: Same resource used in different papers (LREC2014).

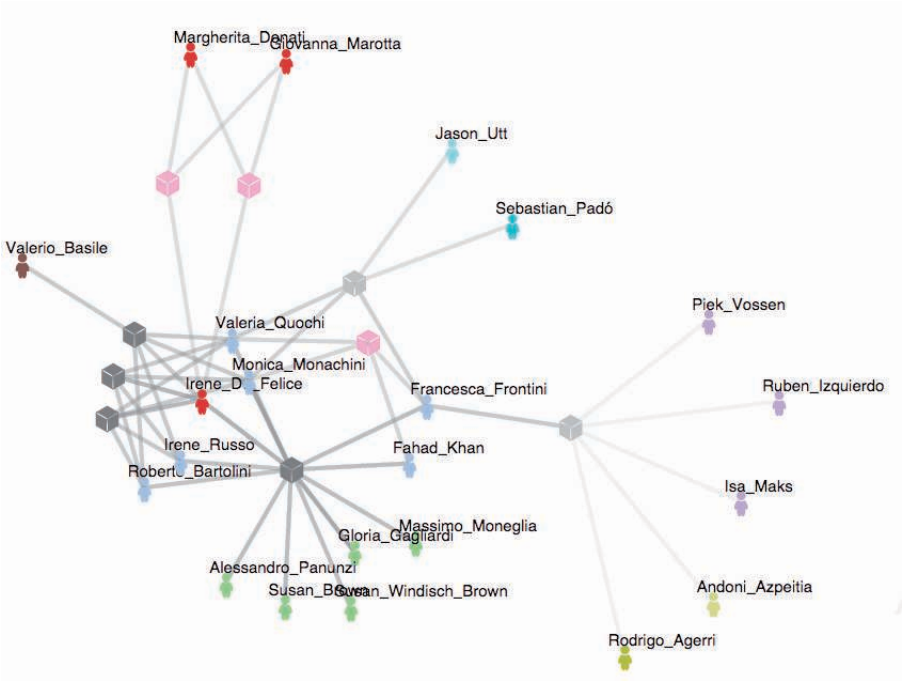


Figure 4: Both conferences together.