LREC as a Graph: People and Resources in a Network

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

Abstract

This proposal describes a new way to visualise resources in the LREMap, a community-built repository of language resource descriptions and uses. The LREMap is represented as a force-directed graph, where resources, papers and authors are nodes. The analysis of the visual representation of the underlying graph is used to study how the community gathers around LRs and how LRs are used in research.

Keywords: language resources, resources documentation, data visualisation

1. Introduction

The availability of Language Resources (LRs) - such as corpora, computational lexicons and parsers - is crucial to most NLP technologies. 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), (Rehm and Uszkoreit, 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 other languages 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, production of quality LRs is only the first step; in order to be usable, 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 Language Resource 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, and 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 terms of validation, documentation and clearing of Intellectual Property Rights (IPR) issues. 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 LREMap was launched as an initiative at LREC 2010 in order to crowdsource reliable and accurate documentation for the largest possible set of resources. Authors 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 initiatives 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 research infrastructure. As the Italian CLARIN representative, 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.

In previous works (Del Gratta et al., 2015) we proposed an analysis of the Italian LR panorama, comparing data drawn from the LRE Map with data manually extracted from the CLiC-It 2014 proceedings. Here we extend our analysis to all languages, using data from LREC 2010, LREC 2012 and LREC 2014 proceedings. In doing this, we build upon previous work by Mariani e Francopoulo (Mariani and Francopoulo, 2015), where data from the LREMap is used to produce Language Matrices “presenting the number of resources of various types that exist for various modalities for each language”, as well as the number of times each resource is mentioned in a paper. For the purpose, they introduce the idea of using bibliometry to evaluate the impact of a language resource, just as it is done for papers or journals, and of calculating a “Language Resource Impact Factor” (LRIF) based on LR mentions in papers. In this paper we shall also attempt to identify LRs that seem to be more or less central to the scientific community network and its research production, and measure the impact of a LR on the research outcomes.In order to do this, we are taking into account the data gathered during various editions of LREC, namely 2010, 2012, 2014, contributed by authors of the main conference. Part of these data (2014) are also available as Linked Data in RDF[2].

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

[2] See http://datahub.io/dataset/lremap-conf and (Del Gratta et al., 2014a).

	#resources	#papers	#people
LREC2010	1177	578	1651
LREC2012	610	398	1331
LREC2014	675	477	1714

Table 1: Data about different LREC editions.

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 (such as *corpus, lexicon, tool*), modality, language(s), use, and availability. To this core set of metadata, the LREMap 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 LREMap. This also means that any given resource can have more than just one entry in the LREMap, 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: for instance, some of the most used resources according to the LREMap are Princeton WordNet and the British National Corpus (BNC).

For the purposes of this paper we only took into consideration the following metadata for each entry in the LREMap: resource name, resource type, authors and affiliations. Basic statistics about the three editions of LREC under analysis are reported in Table 1: even if there has been a decrease since 2010 in the input about resources provided by authors, numbers are still interesting, especially when enriched with information about co-authorship for the visualisation of social networks graphs (see Table 3.). The analysis of co-authorship networks in the field of computational linguistics is not new: thanks to the ACL ANTHOLOGY NETWORK initiative (Radev et al., 2009) bibliographic data about papers' citations and authors' collaboration from the ACL Anthology are easy to explore [3]. In a similar vein Saffron [4] (Buitelaar et al., 2014; Bordea et al., 2013; Buitelaar et al., 2013) as a research framework based on text mining and linked data principles is able to perform community detection suggesting domain specific experts. Visualisations are organised around topics automatically extracted.

The kind of networks we analyse in this paper are focused on the building blocks of scientific work in the field of computational linguistics, language resources. People can be connected because they jointly worked to write a paper but more significantly they can be connected because they used the same resource or resources with similar features (for example concerning the same type or the same language). They can be close in terms of interests and past experiences with tools, corpora, lexicon etc. Discovering this in a graphical form can enhance the awareness of the role of language resources and the desire to document them properly to highlight how much interconnected is each piece of work in the scientific community. As a matter of fact a scientific community where each author writes one paper working with just one language resource is not promoting exchange of ideas and is not promising for theoretical and practical improvements. With this assumption in mind we report frequency data (both absolute and relative) in Table 2 and Table 3 concerning four types of resources documented by authors of LREC 2010, LREC 2012 and LREC 2014.

	res10	res12	res14
Lexicon	115 (20%)	58 (14,6%)	62 (13%)
Corpus	334 (57,8%)	234 (58,8%)	278 (58,3%)
Annotation Tool	67 (5,6%)	38 (6,2%)	28 (4,2%)
Tagger/ Parser	99 (8,5%)	31 (2,8%)	22 (3,3%)

Table 2: Frequencies of four types of resources in different editions of LREC.

With the idea of understanding how interconnected the community is around a resource type and whether data from different editions of the same conference helps in the detection of trends we consider:

- the ratio between the number of distinct authors in Table 3 (value *auth*) and the number of resources for every type in analysis in Table 2 (value *res*);
- the ratio between the number of papers in Table 3 (value *paper*) and the number of resources for every type in analysis in Table 2 (value *res*).

The worst case (not occurring in our dataset) happens when the number of papers is equal to the number of resources: there are not papers available comparing/using more resource of the same type. Similarly, when the ratio between authors' number and resources' number is close to 1 means that people didn't work together with the same resource/tool. As an indicator of social network richness and complexity, the ratio between the two values listed above help us to discover a subset of data that produce an interesting visualisation. For example, resources of type *Lexicon* at LREC2014 produces a wider and more interconnected graph with respect to resources of type *Annotation Tool* at the same edition. Even if the aim of this paper is suggesting visualisation as a mean to explore and to understand data about language resources we verified its feasibility with these preliminary analyses.

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).

[3] http://clair.eecs.umich.edu/

[4] link: http://saffron.deri.ie/

	auth10	auth12	auth14	papers10	papers12	papers14
Lexicon	169 (10%)	118 (8,9%)	175 (10%)	115 (20%)	58 (14,6%)	62 (13%)
Corpus	899 (54%)	717 (53%)	997 (59%)	334 (57,8%)	234 (58,8%)	278 (58,3%)
Annotation Tool	91 (5,6%)	90 (6,8%)	58 (3,4%)	60 (10,4%)	35 (8,8%)	28 (5,9%)
Tagger/Parser	116 (4,8%)	34 (2,6%)	60 (3,5%)	67 (11,6%)	21 (5,3%)	21 (4,5%)

Table 3: Frequencies of authors and papers associated with four types of resources in different editions of LREC.

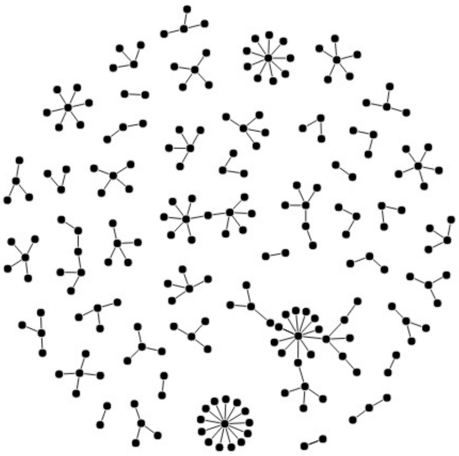

Figure 1: Snapshot of authors and resources of type *Lexicon* at LREC2014.

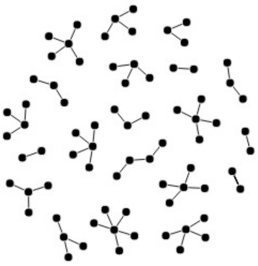

Figure 2: Snapshot of authors and resources of type *Annotation Tool* at LREC2014.

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 three conferences (LREC 2010, LREC 2012 and LREC 2014), the aim was to portray the 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 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 4). 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.

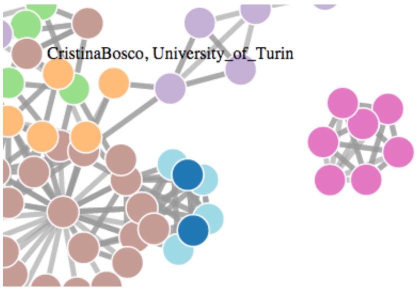

Figure 3: Snapshot of the first graph. Co-authorship clusters.

4. Conclusions and Future Work

In this work we use visualisations to show how the NLP community uses LRs in the works presented at three editions of LREC (LREC2010, LREC2012 and LREC2014). We highlight how collaborations cluster around the use of major resources, and how networks are created by users of the same resources. This analysis is part of the activity that CNR-ILC, as a CLARIN node, will actively promote. We will help the LR community (both creators and users) improve the documentation of LRs, thus making them more widely known to others and ensuring 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

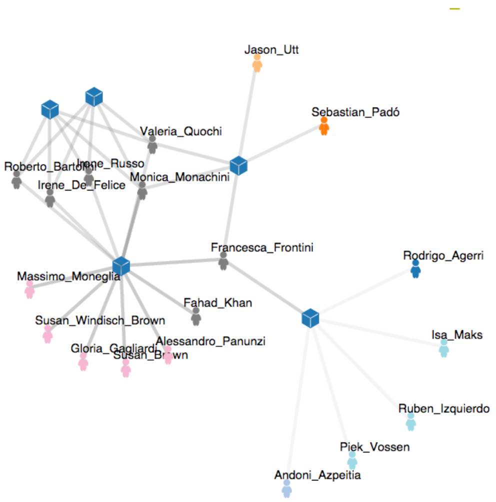

Figure 4: Snapshot of the second graph. Authors connected via resources.

well as the porting of LR descriptions to emerging channels and formats (LINGhub[5], RDF-LOD).

5. Acknowledgements

The research presented in this paper was carried out within the framework of CLARIN-IT and was partly funded by SM@RTINFRA (MIUR Progetto premiale) and PARTHENOS (H2020 INFRADEV-4).

6. Bibliographical References

Bordea, G., Bogers, T., and Buitelaar, P. (2013). Benchmarking domain-specific expert search using workshop program committees. In *Proceedings of the 2013 workshop on Computational scientometrics: theory & applications*, pages 19–24. ACM.

Broeder, D., Kemps-Snijders, M., Van Uytvanck, D., Windhouwer, M., Withers, P., Wittenburg, P., and Zinn, C. (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), Paris.

Buitelaar, P., Bordea, G., and Polajnar, T. (2013). Domain-independent term extraction through domain modelling. In *The 10th International Conference on Terminology and Artificial Intelligence (TIA 2013), Paris, France.* 10th International Conference on Terminology and Artificial Intelligence.

Buitelaar, P., Bordea, G., and Coughlan, B. (2014). Hot topics and schisms in NLP: community and trend analysis with saffron on ACL and LREC proceedings. In Nicoletta Calzolari, et al., editors, *Proceedings of the Ninth International Conference on Language Resources xand Evaluation (LREC'14), Reykjavik, Iceland, May 26-31, 2014.*, pages 2083–2088. European Language Resources Association (ELRA), Paris.

Cairo, A. (2013). *L'arte funzionale: Infografica e visualizzazione delle informazioni*. Pearson Italia Spa.

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 *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 1084–1089. European Language Resources Association (ELRA), Paris.

Del Gratta, R., Frontini, F., Khan, F., Mariani, J., and Soria, C. (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*.

Del Gratta, R., Goggi, S., and Pardelli, G. (2014b). LRE Map disclosed. In *Proceedings of the ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 3534–3541. European Language Resources Association (ELRA), Paris.

Del Gratta, R., Frontini, F., Monachini, M., Pardelli, G., Russo, I., Bartolini, R., Goggi, S., Khan, F., Quochi, V., Soria, C., and Calzolari, N. (2015). Visualising Italian Language Resources: a Snapshot. In *Proceedings of the Second Italian Conference on Computational Linguistics CLiC-it 2015*, pages 100–104.

Gavrilidou, M., Labropoulou, P., Desipri, E., Piperidis, S., Papageorgiou, H., Monachini, M., Frontini, F., Declerck, T., Francopoulo, G., Arranz, V., 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), Paris.

Mariani, J. and Francopoulo, G. (2015). Language matrices and a language resource impact factor. In *Language Production, Cognition, and the Lexicon*, pages 441–471. Springer, Netherlands.

McCrae, J., Labropoulou, P., Gracia, J., Villegas, M., Rodriguez-Doncel, V., and Cimiano, P. (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*.

Radev, D. R., Joseph, M. T., Gibson, B., and Muthukrishnan, P. (2009). A Bibliometric and Network Analysis of the field of Computational Linguistics. *Journal of the American Society for Information Science and Technology*.

G. Rehm et al., editors. (2012). *META-NET White Paper Series: Europe's Languages in the Digital Age*. Springer.

Soria, C., Bel, N., Choukri, K., Mariani, J., Monachini, M., Odijk, J., Piperidis, S., Quochi, V., Calzolari, N., 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), Paris.

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