Editorial of the Special Issue on Latest Advancements in Linguistic Linked Data

Julia Bosque-Gil [a,*], Philipp Cimiano [b] and Milan Dojchinovski [c]

[a] *Aragon Institute of Engineering Research (I3A), University of Zaragoza, Spain*
E-mail: jbosque@unizar.es
[b] *Cognitive Interaction Technology Center, Bielefeld University, Germany*
E-mail: cimiano@cit-ec.uni-bielefeld.de
[c] *Faculty of Information Technology, Czech Technical University in Prague, Czech Republic*
E-mail: milan.dojchinovski@fit.cvut.cz

Editors: Pascal Hitzler, Kansas State University, USA; Krzystof Janowicz, University of California, Santa Barbara, USA

Abstract. Since the inception of the Open Linguistics Working Group in 2010, there have been numerous efforts in transforming language resources into Linked Data. The research field of Linguistic Linked Data (LLD) has gained in importance, visibility and impact, with the Linguistic Linked Open Data (LLOD) cloud gathering nowadays over 200 resources. With this increasing growth, new challenges have emerged concerning particular domain and task applications, quality dimensions, and linguistic features to take into account. This special issue aims to review and summarize the progress and status of LLD research in recent years, as well as to offer an understanding of the challenges ahead of the field for the years to come. The papers in this issue indicate that there are still aspects to address for a wider community adoption of LLD, as well as a lack of resources for specific tasks and (interdisciplinary) domains. Likewise, the integration of LLD resources into Natural Language Processing (NLP) architectures and the search for long-term infrastructure solutions to host LLD resources continue to be essential points to which to attend in the foreseeable future of the research line.

1. Introduction

Linguistic Linked Data (LLD) refers to the application of linked data principles to the representation and publication of linguistic resources, including among them lexica, dictionaries, corpora, terminologies, metadata repositories and linguistic ontologies. As is the case with data from other domains, the linked data paradigm in this setting allows to share language resources in a uniform and interoperable way that facilitates and enhances their discovery, integration and reuse. One of the communities acting as main driver in this endeavour has been the *Open Linguistics Working Group* (OWLG) [1, 2] gathered in the context of the Open Knowledge Foundation (OKFN)[1].

The work in this community, together with other European projects and initiatives (e.g. LIDER[2]) in the past decade has led to the creation of the Linguistic Linked Open Data (LLOD) cloud, an LOD subset comprising of linguistic resources [1, 3][3]. In 2016, the LLOD cloud had already experienced a growth by a factor of 4 since its first

[*] Corresponding author. E-mail: jbosque@unizar.es.
[1] https://okfn.org/
[2] https://lider-project.eu
[3] https://linguistic-lod.org/llod-cloud

instantiation [1], and was growing at 19.3% per year in the period 2018-2020 [2]. Nowadays, it amounts to more than 200 resources (on the publication of these LLOD cloud resources and their metadata, see di Buono et al., this volume).

This continuous growth is to a large extent due to the efforts coming from projects and initiatives in the fields of computational linguistics, computer science, information technology, lexicography, and applied linguistics. H2020 projects such as Lynx[4], Prêt-à-LLOD[5], ELEXIS[6], and COST Actions such as ENEL[7] or NexusLinguarum[8] include linguistic linked data as a key technology to build common infrastructures and domain-specific knowledge graphs, to grow networks around the emerging field of linguistic data science, and to establish robust ecosystems to address the life-cycle of language resources: from identifying the requirements concerning the representation of linguistic content to its exploitation by natural language processing (NLP) applications in a variety of sectors.

In its initial stage, the linguistic linked data line of research motivated the development and/or conversion of multiple LLD datasets and vocabularies to account for the representation needs of different types of resources. Models such as the Lexicon Model for Ontologies (*lemon*) [4] (and its successor OntoLex [5]), the NLP Interchange Format (NIF) [6], the Meta-Share Ontology [7] or the Ontologies for Linguistic Annotation (OLiA) [8] emerged. Likewise, datasets commonly used as a linking backbone were represented as linked data or newly developed (WordNet [9, 10], BabelNet [11]), and linguistic linked data categories such as LexInfo [12] started to draw the attention of the community and slowly became a *de facto* standard to encode morphosyntactic categories. As of today, there are multiple vocabularies (see Fahad Khan et al. in this volume) and resources providing numerous options to link to relevant and supplementary data enriching the content of a given language resource.

With the rapid growth and the increasing interest in the use of linked data for NLP in this decade, new challenges have emerged concerning particular use cases, domain and task applications, quality dimensions, and linguistic features for which to account. Some of these aspects refer to the application of linked data principles in general and are not particularly tied to the linguistics domain, whereas others are concerned with the nature of the linguistic content of the resource.

2. Motivation and overview of the special issue

Since the inception of the Open Linguistics Working Group in 2010, there have been many efforts in transforming language resources into Linked Data. The research field of Linguistic Linked Data (LLD) has gained in importance, visibility and impact. The editors of this special issue have thus decided to publish a special issue on latest advancements in the LLD with two main goals:

- reviewing and summarizing the progress and status of LLD research in recent years
- developing an understanding of the challenges ahead of the field for the years to come

The call for papers for the special issue was distributed over relevant mailing lists. As a response to our call, we received 13 relevant submissions, of which after a thorough reviewing process involving three reviewers for each paper at least, and at least 2 rounds of improvements, we finally decided to accept eight papers.

This issue includes papers on a range of topics such as models for linguistic data representation, metadata and its quality assessment, diachronicity, typology, and, closer to the NLP field, entity linking, terminology extraction, bilingual dictionary generation, and the detection of discourse relations.

The paper by Kahn et al. provides an overview of the current state of ontologies and vocabularies used to describe linguistic resources as linked data. The main emphasis lies on understanding how these models can support the FAIR publication of language resources. The authors provide an overview over the main vocabularies for describing corpora (NIF, Open Annotation), lexica and dictionaries (OntoLex Lemon, SKOS), terminologies (MMOn

[4] https://lynx-project.eu/
[5] https://pret-a-llod.github.io/
[6] https://elex.is/
[7] https://www.cost.eu/actions/IS1305/
[8] https://nexuslinguarum.eu/

ontology and PHOIBLE) as well as linguistic resource metadata (METASHARE Ontology, ISOcat, GOLD) in addition to vocabularies for describing typological datasets. The authors include in particular a summary of the main recent developments of a selection of these models. Further, the paper provides an overview of projects that are currently contributing to the growth and development of standards for the LLOD. As main challenges for the future, the authors emphasize the need for reliable infrastructure that supports the longer term hosting and accessibility of resources. Second, the authors highlight that as the complexity of the models developed for the description of language resources as LLD increases, one challenge is to support the consistent and correct use of the vocabularies by end users (engineers, linguists, etc). For this, ontology design patterns and templates that describe best (modelling) practices would be an important asset.

The paper by Pia di Buono is concerned with analyzing the current state and adoption of standards for the description of metadata. They analyse the LOD Cloud and Annohub[9] in terms of the availability of metadata. For this, they develop a manual mapping of metadata properties in these resources into the METASHARE schema. The focus of their analysis of the metadata is on the provision of information about domain, language, and license information. Further, an important focus is resource accessibility via data dump or SPARQL endpoint. The authors find that most of the resources available on the LLOD are corpora, lexicons and dictionaries. The most frequent languages are in addition to English Swedish, Spanish, German, French and Italian. The unique identification of languages continues to be problematic as resources use different codes. Regarding licenses, most resources use open licenses (CC-BY variants or GLP). Regarding accessibility, about 70% provide a data dump and less than 30% have a working SPARQL endpoint. This corroborates the need for better and efficient solutions for long-term hosting of data as identified by Kahn et al. (see above). In addition to their analysis, Pia di Buono et al. propose a new scheme and resource (MELLD) that is the result of mapping the metadata schema of the LLOD cloud and Annohub to the METASHARE model. It remains an open question to see if this model might be the foundation for a uniform metadata standard for LLD resources.

In their contribution "Glottocodes: Identifiers Linking Families, Languages and Dialects to Comprehensive Reference Information", Hammarstrom and Forkel present Glottocodes, an identification system for *languoids* (languages, dialects and language families) in Glottolog. Glottolog gathers language data and supporting bibliographic references, grouping data into specific levels (e.g. dialect, language, subfamily, etc). This categorisation, however, might change throughout time, given the controversies on language vs. dialect status or as more knowledge on a particular languoid is recorded. The authors propose Glottocodes as a system of persistent identifiers for languoids which improves on the ISO 639-3 language identifiers. The resource has been conceived to support machine readability and stays independent of the level of linguistic abstraction (idiolect, dialect, language, family, etc.) and its potential changes. As such, one of the key challenges addressed by their proposal is the dynamic nature of languoids and hence the need for technological solutions to consider the continuous changes and updates. The authors also address the topic of an optimal infrastructure to facilitate data curation and the involvement of the wider community. In the case of Glottolog, the authors turned to `git` and GitHub for hosting and version control tracking as means to facilitate the contribution, edition, or request submission by the community. To publish the released versions, as well as for for archiving and indexing purposes, Zenodo is the repository used. In addition, the authors provide a detailed assessment of the resulting resource, Glottolog, in terms of FAIR principles.

Armaselu et. al present a thorough survey addressing jointly the detection of semantic change in multilingual diachronic corpora in NLP and its representation as LLOD. The study focuses on the generation of diachronic ontologies, and aims to provide a first step towards bridging the gap between NLP and LLOD from the perspective of humanities research. To do so, the authors propose a workflow for this interdisciplinary study, revisiting the works on semantic change from different theoretical frameworks, its analysis and representation in the Semantic Web context, its detection with different NLP approaches and tools, and lastly the generation of diachronic linked data resources and their subsequent publication. A major challenge faced by the authors concerns the interdisciplinary nature of the study itself, involving lines of work with different maturity levels. In relation to this, the need for a framework to foster collaboration, exchanges and communication across the various research lines is highlighted. The limitations in representing temporal and dynamic information as LLOD and the absence of guidelines to generate diachronic

[9] https://annohub.linguistik.de/de/

ontologies represent significant barriers for the adoption of LLOD. Finally, and common to other lines of work involving the humanities community and LLOD researchers, the need for methods to facilitate the publication and maintenance of linked data for non-Semantic Web experts is identified by the authors.

Özel et al. address the scarcity of multilingual resources for discourse analysis. A major contribution of their work are the two methods for discourse relation linking in the TED Multilingual Discourse Treebank (TED-MDB), one based on word alignment, and a second one on cross-lingual sentence embeddings. The TED-MDB is annotated with discourse relations between sentences in six different languages. However, the annotations on the different languages were performed independently from each other, which hinders the cross-lingual analysis of discourse connectives and motivates the authors' relation linking task. Their results show that the cross-lingual sentence embeddings approach outperforms the word alignment one, which is negatively affected by incorrectly derived sentence alignments. Thanks to the extracted relations, the authors gain new insights into discourse structures present in the TED-MDB, and generate bilingual discourse connectivity lexica relevant for machine translation, discourse studies and language teaching. The discourse relation linking task is still a challenging problem to tackle due to argument spans of relations varying across languages or multiple relations overlapping. In a broader sense, the scarcity of multilingual discourse resources (not necessarily linked data-based) remains low, although this work makes a step forward to increase their availability.

In their "Survey on English Entity Linking on Wikidata", Cedric Moeller, Jens Lehmann and Ricardo Usbeck present the results from a survey on Entity Linking datasets and approaches in the context of Wikidata. The authors argue that the vast majority of Entity Linking approaches consider specific properties such as labels and descriptions, however the contextual information, the structure and links of Wikidata, is rarely exploited. Furthermore, the survey reveals that most of the Entity Linking datasets are created as "mapped version" of already existing datasets, i.e. not exclusively focused on Wikidata, and the time-variance and multilingualism aspects are still poorly represented.

Over the last decade, large number of dictionaries have been converted, linked and published as part of the LLOD cloud. The availability of linked dictionaries provides new opportunities for their exploitation. In the paper "Bilingual dictionary generation and enrichment via graph exploration", Shashwat Goel, Jorge Gracia and Mikel L. Forcada propose a novel method that exploits the graph structure of existing bilingual linked dictionaries and infers new bilingual entries. The method has been applied and validated on the Apertium knowledge graph which produced new bilingual dictionaries with 70% the size of the source Apertium dictionaries at a precision of 85%.

Multilingual terminologies play an important role in many language technology solutions. Their creation typically requires significant amount of human effort and due to their availability in different formats their reuse is limited. In "TermitUp: Generation and Enrichment of Linked Terminologies", Patricia Martín-Chozas, Karen Vázquez-Flores, Pablo Calleja, Elena Montiel-Ponsoda and Víctor Rodríguez-Doncel present TermitUp, a service for automated extraction of domain-specific terminologies and their enrichment with data from the Linguistic LOD cloud. The created terminologies are validated, linked with other terminological resources and published as part of the Linguistic LOD cloud.

3. Discussion

The papers in the special issue clearly convey that the field of linguistic linked data (LLD) has certainly progressed and matured over the years. The field has seen a convergence in terms of ontologies / models used for the description of data and metadata and best practices have been identified. Yet, the special issue also shows that there are important challenges to address in the future.

Adoption. It is still difficult to discover and reuse LLD. As identified by Maria Pia di Buono et al. interoperable metadata standards are not yet fully available and used. This is clearly an obstacle for the use of LLD datasets which represents a significant barrier to the practical adoption. The paper by Armaselu et al. points out an important challenge regarding adoption, too, emerging from the difficulties of advancing in interdisciplinary approaches with lines of varying level of maturity, and the need to set up frameworks to facilitate exchange and collaboration across fields. In this regard, bringing LLD generation, publication and maintenance closer to non-experts in Semantic Web remains an important line of research and dissemination.

Representation needs. Although the availability of LLD-related vocabularies and their coverage has significantly increased in the past years, further work and best practices are needed to address the representation needs of linguistic data relevant for areas under-represented in the LLOD cloud (e.g. as opposed to synchronic lexical semantics). This is the case of diachronic information, as pointed out by Armaselu et. al. In relation to language as the object of study as well, the persistent identifiers proposed by Hammarstrom and Forkel (Glottocodes) to remain agnostic to changes in the description of languages, language families, dialects, etc. in Glottolog serve as a reminder of the need of solutions that adapt to potential updates as more information on a given language is obtained, which is particularly relevant for languages under-resourced as of today.

Scarcity of resources. Closely tied to the previous point, the low number of cross-lingual discourse resources available pointed out by Özel et. al. indicates that further work on that line would also lead to a greater balance in the availability of potential LLD resources and models covering different linguistic levels. This scarcity also holds for the availability of datasets for particular NLP tasks such as Entity Linking. As identified by Moeller et al. there is a lack of Entity Linking evaluation datasets which consider multilingualism and time-variance.

Integration into NLP architectures. A further challenge is to facilitate the integration of linguistic linked dataset into NLP architectures and systems so that systems can be easily ported to work on a new dataset. Again this requires the consistent use of vocabularies and standards to describe the language resources at the content level, so that resources become directly pluggable into workflow.

Infrastructure. The need for a reliable infrastructure that supports the longer term hosting and accessibility of resources has been called to attention by Kahn et al. as well as by other works in the recent literature [13]. A challenge for the future is to identify requirements regarding the availability of LLD resources and develop (business) models to cover the costs incurred by the services required to maintain availability. A further challenge lies in creating incentives to foster involvement of the whole community in the curation and enrichment of data.

However, solutions that guarantee the availability of LLD resources in long-term are still to be thoroughly discussed and addressed as a key aspect for the future development of the LLOD cloud and the line of research as a whole. The involvement of the wider community in data curation in feedback or request gathering through a suitable infrastructure is also a significant point to bear in mind, with suggested solutions as the one presented by Hammarstrom and Forkel.

The Linguistic LOD cloud has been under development for over a decade now and we already see some works which exploit its potential, such as the work on bilingual dictionary creation by Shashwat Goel et al., and the domain specific terminology extraction by Patricia Martin-Chozas et al. However, the potential of the LLOD cloud is enormous and we expect to see many more works in the near future which will exploit the datasets available as part of the LLOD cloud.

References

[1] J. McCrae, C. Chiarcos, F. Bond, P. Cimiano and T. Declerck, The Open Linguistics Working Group: Developing the Linguistic Linked Open Data Cloud, in: *Proceedings of 10th Language Resources and Evaluation Conference (LREC 2016)*, N. Calzolari, K. Choukri, T. Declerck, S. Goggi, M. Grobelnik, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk and S. Piperidis, eds, European Language Resources Association (ELRA), Paris, France, 2016, pp. 2435–2441. http://iiis.tsinghua.edu.cn/~weblt/papers/llod-lrec2016.pdf.

[2] C. Chiarcos, B. Klimek, C. Fäth, T. Declerck and J.P. McCrae, On the Linguistic Linked Open Data Infrastructure, in: *Proceedings of the 1st International Workshop on Language Technology Platforms*, G. Rehm, K. Bontcheva, K. Choukri, J. Hajič, S. Piperidis and A. Vasiljevs, eds, European Language Resources Association, Marseille, France, 2020, pp. 8–15. ISBN 979-10-95546-64-1. https://aclanthology.org/2020.iwltp-1.2.

[3] C. Chiarcos, S. Hellmann and S. Nordhoff, Towards a Linguistic Linked Open Data Cloud: The Open Linguistics Working Group, *TAL Traitement Automatique des Langues* (2011), 245–275. ISBN 978-3-642-31781-1. doi:10.1007/978-3-642-31782-8.

[4] J. McCrae, G. Aguado-de-Cea, P. Buitelaar, P. Cimiano, T. Declerck, A. Gómez-Pérez, J. Gracia, L. Hollink, E. Montiel-Ponsoda, D. Spohr and T. Wunner, Interchanging lexical resources on the Semantic Web, *Language Resources and Evaluation* **46** (2012), 701–719. ISBN 1574-020X. doi:10.1007/s10579-012-9182-3.

[5] J.P. McCrae, J. Bosque-Gil, J. Gracia, P. Buitelaar and P. Cimiano, The Ontolex-Lemon model: development and applications, in: *Proceedings of eLex 2017 conference*, I. Kosem, C. Tiberius, M. Jakubíček, J. Kallas, S. Krek and V. Baisa, eds, Lexical Computing CZ s.r.o, Brno, Czech Republic, 2017, pp. 19–21.

[6] S. Hellmann, J. Lehmann, S. Auer and M. Brümmer, Integrating NLP using linked data, in: *International Semantic Web Conference*, H. Alani, L. Kagal, A. Fokoue, P. Groth, C. Biemann, J.X. Parreira, L. Aroyo, N. Noy, C. Welty and K. Janowicz, eds, Springer, Berlin, Heidelberg, 2013, pp. 98–113.

[7] J.P. McCrae, P. Labropoulou, J. Gracia, M. Villegas, V. Rodriguez-Doncel and P. Cimiano, One ontology to bind them all: The META-SHARE OWL ontology for the interoperability of linguistic datasets on the Web, in: *The Semantic Web. Latest Advances and New Domains. 12th European Semantic Web Conference*, F. Gandon, M. Sabou, H. Sack, C. d'Amato, P. Cudré-Mauroux and A. Zimmermann, eds, Lecture Notes in Computer Science, Vol. 9088, Springer, 2015, pp. 271–282. ISBN 978-3-319-18817-1.

[8] C. Chiarcos and M. Sukhareva, Olia–Ontologies of Linguistic Annotation, *Semantic Web* **6**(4) (2015), 379–386.

[9] M. van Assem, A. Gangemi and G. Schreiber, Conversion of WordNet to a standard RDF/OWL representation, in: *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*, N. Calzolari, K. Choukri, A. Gangemi, B. Maegaard, J. Mariani, J. Odijk and D. Tapias, eds, European Language Resources Association (ELRA), Genoa, Italy, 2006. http://www.lrec-conf.org/proceedings/lrec2006/pdf/165_pdf.pdf.

[10] J. McCrae, C. Fellbaum and P. Cimiano, Publishing and Linking WordNet using lemon and RDF, in: *Proceedings of the 3rd Workshop on Linked Data in Linguistics*, C. Chiarcos, J.P. McCrae, P. Osenova and C. Vertan, eds, Association for Computational Linguistics, Stroudsburg, Pennsylvania, 2014.

[11] R. Navigli and S.P. Ponzetto, BabelNet: The Automatic Construction, Evaluation and Application of a Wide-Coverage Multilingual Semantic Network, *Artificial Intelligence* **193** (2012), 217–250.

[12] P. Cimiano, P. Buitelaar, J. McCrae and M. Sintek, LexInfo: A declarative model for the lexicon-ontology interface, *Journal of Web Semantics* **9**(1) (2011), 29–51.

[13] C. Chiarcos, Get! Mimetypes! Right!, in: *3rd Conference on Language, Data and Knowledge (LDK 2021)*, D. Gromann, G. Sérasset, T. Declerck, J.P. McCrae, J. Gracia, J. Bosque-Gil, F. Bobillo and B. Heinisch, eds, Open Access Series in Informatics (OASIcs), Vol. 93, Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl, Germany, 2021, pp. 5:1–5:4. ISBN 978-3-95977-199-3. doi:10.4230/OASIcs.LDK.2021.5. https://drops.dagstuhl.de/opus/volltexte/2021/14541.