information

MDPI

Editorial

Foreword to the Special Issue: "Towards the Multilingual Web of Data"

John P. McCrae [1,*] and Jorge Gracia [2,*]

[1] Data Science Institute/Insight Centre for Data Analytics, National University of Ireland Galway, H91 CF50 Galway, Ireland

[2] Aragon Institute of Engineering Research (I3A), University of Zaragoza, 50018, Zaragoza, Spain

* Correspondence: john@mccr.ae (J.P.M.); jogracia@unizar.es (J.G.)

Received: 7 February 2019; Accepted: 7 February 2019; Published: 9 February 2019

We are pleased to introduce this special issue on the topic of "Towards the Multilingual Web of Data", which we feel is a timely and valuable topic in our increasingly multilingual and interconnected world. The Web of Data has increasingly become a space where concepts are described not only with logic and ontologies but also with linguistic information in the form of multilingual lexicons, terminologies and thesauri. In particular, this has led to the creation of a growing cloud of linguistic linked open data, which bridges the world of ontologies with dictionaries, corpora and other linguistic resources. This raises several challenges, such as ontology localization, cross-lingual question answering, cross-lingual ontology and data matching, representation of lexical information on the Web of Data, etc.

Furthermore, Natural Language Processing (NLP) and machine learning for linked data can benefit from exploiting multilingual language resources, such as annotated corpora, wordnets, bilingual dictionaries, etc., if they are themselves formally represented and linked by following the linked data principles. A critical mass of language resources as linked data on the Web are leading to a new generation of linked data-aware NLP techniques and tools which, in turn, will serve as basis for a richer, multilingual Web.

In this special issue, we are pleased to publish six high-quality publications related to the topic of the multilingual Web of Data and in particular, with relation to models that are widely used in the Web already, such as the OntoLex-Lemon model. Four of them were extended versions of the best papers published at the 6th Workshop on Linked Data in Linguistics (LDL-2018) that took place in Miyazaki, Japan on 12 May 2018, co-located with the Language Resources and Evaluation Conference (LREC) 2018. The other two papers were regular submissions to the Special Issue.

We are also pleased to see the breadth of the coverage of topics with three papers focusing on issues related to minority and historical languages, showing the increasing diversity of datasets handled by modern technologies. In addition, we have methods presented that are related to language-agnostic methods, generic representation of data and data collection.

In "Annotating a Low-Resource Language with LLOD Technology: Sumerian Morphology and Syntax" by Christian Chiarcos, Ilya Khait, Émilie Pagé-Perron, Niko Schenk, Jayanth, Christian Fäth, Julius Steuer, William Mcgrath and Jinyan Wang [1], we see that the authors present a methodology for representing some of the earliest known texts in Sumerian, focusing on how modern Web standards, such as SPARQL in combination with novel representations such as CoNLL-RDF, can make exploring such corpora much easier. Similarly, in "Multilingual and Multiword Phenomena in a Lemon Old Occitan Medico-Botanical Lexicon" by Andrea Bellandi, Emiliano Giovannetti and Anja Weingart [2], the authors explore the development of a lexicon in the historical romance language of Old Occitan, and explore how the use of the OntoLex-Lemon model can help in representing this data, especially in relation to multiword expressions. The use of OntoLex-Lemon is further explored by Frances Gillis-Webber in "Conversion of the English-Xhosa Dictionary for Nurses to a Linguistic Linked

Data Framework" [3], where the conversion of a bilingual dictionary from English to Xhosa, a Bantu language of South Africa, into linked data is explored. The challenges of the data representation and the publishing of the resource in a framework is explored and would be useful for those who need to publish dictionaries on the Web. In "Towards the Representation of Etymological Data on the Semantic Web" [4], Anas Fahad Khan proposes a new extension to the OntoLex-Lemon model for the representation of etymology information in the context of dictionaries. This new model considers the challenge of representing the different processes by which historic forms of words have developed and been borrowed into modern languages and the challenges in representing them.

While lexicography and minority languages are a strong focus of this special issue, language-agnostic methods are also important in this context, such as those explored by Nicolas Heist, Sven Hertling and Heiko Paulheim in "Language-Agnostic Relation Extraction from Abstracts in Wikis" [5], where they extracted 1.6 million new relations from Wikipedia, primarily by exploiting background knowledge from the graph. As such, this method is highly applicable to the approximately 300 languages for which a Wikipedia has been created. Finally, in "Semantic Modelling and Publishing of Traditional Data Collection Questionnaires and Answers" [6] by Yalemisew Abgaz, Amelie Dorn, Barbara Piringer, Eveline Wandl-Vogt and Andy Way, the authors consider the collection and publication as linked data of a collection of questionnaires describing the Bavarian dialects of German in Austria and show how linked data can make this data more useful.

These papers provide an interesting and deep analysis of the challenges of using modern technology in the representation of data with application to linguistics and NLP and present an excellent overview of the state-of-the-art in this field. We hope that you enjoy reading these articles as much as we do.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Chiarcos, C.; Khait, I.; Pagé-Perron, É.; Schenk, N.; Jayanth; Fäth, C.; Steuer, J.; Mcgrath, W.; Wang, J. Annotating a Low-Resource Language with LLOD Technology: Sumerian Morphology and Syntax. *Information* **2018**, *9*, 290. [CrossRef]
2. Bellandi, A.; Giovannetti, E.; Weingart, A. Multilingual and Multiword Phenomena in a *lemon* Old Occitan Medico-Botanical Lexicon. *Information* **2018**, *9*, 52. [CrossRef]
3. Gillis-Webber, F. Conversion of the *English-Xhosa Dictionary for Nurses* to a Linguistic Linked Data Framework. *Information* **2018**, *9*, 274. [CrossRef]
4. Khan, A.F. Towards the Representation of Etymological Data on the Semantic Web. *Information* **2018**, *9*, 304. [CrossRef]
5. Heist, N.; Hertling, S.; Paulheim, H. Language-Agnostic Relation Extraction from Abstracts in Wikis. *Information* **2018**, *9*, 75. [CrossRef]
6. Abgaz, Y.; Dorn, A.; Piringer, B.; Wandl-Vogt, E.; Way, A. Semantic Modelling and Publishing of Traditional Data Collection Questionnaires and Answers. *Information* **2018**, *9*, 297. [CrossRef]