Linked Data in Linguistics

Christian Chiarcos • Sebastian Nordhoff •
Sebastian Hellmann
Editors

Linked Data in Linguistics

Representing and Connecting Language Data and Language Metadata

Springer

Editors
Christian Chiarcos
Information Sciences Institute
University of Southern California
Marina del Rey, CA, USA

Sebastian Hellmann
Business Information Systems
University of Leipzig
Leipzig, Germany

Sebastian Nordhoff
Department of Linguistics
Max-Planck Institute for
Evolutionary Anthropology
Leipzig, Germany

ISBN 978-3-642-28248-5 e-ISBN 978-3-642-28249-2
DOI 10.1007/978-3-642-28249-2
Springer Heidelberg Dordrecht London New York

Library of Congress Control Number: 2012933118

Preface

The explosion of information technology has led to a substantial growth in quantity, diversity and complexity of web-accessible linguistic data. These resources become even more useful when linked. This volume provides an overview over recent developments, use cases, applications and recommendations for the application of the linked data paradigm to represent, exploit, store, and connect different types of linguistic data collections.

Recent relevant developments include: (1) Language archives for language documentation, with audio, video, and text transcripts from hundreds of (endangered) languages. (2) Typological databases with typological and geographical data about languages from all parts of the globe. (3) Development, distribution and application of lexical-semantic resources in Natural Language Processing. (4) Multi-layer annotations and semantic annotation of corpora by corpus linguists and computational linguists, often accompanied by the interlinking of corpora with lexical-semantic resources.

The general trend of providing data online is accompanied by newly developing possibilities to interconnect linguistic data and metadata. This includes general-purpose knowledge bases such as the DBpedia (a machine-readable edition of the Wikipedia), but also repositories with specific linguistic information about languages, as well as about linguistic categories and phenomena.

It is the challenge of our time to store, interlink and exploit this wealth of data, e.g., by modeling different language resources as Linked Data. The contributions assembled in this volume illustrate the band-width of applications of the Linked Data paradigm for representative types of language resources, including lexical-semantic resources, annotated corpora, typological databases as well as terminology and metadata repositories. The book includes representative applications from different fields, ranging from theoretical linguistics (e.g., typology) over applied linguistics (e.g., language documentation, lexicography, and corpus linguistics) to computational linguistics, Natural Language Processing and information technology.

This volume accompanies the Workshop on Linked Data in Linguistics (LDL-2012), held March 7th-9th, 2012 in Frankfurt a. M., Germany, organized by the

Open Linguistics Working Group (OWLG) of the Open Knowledge Foundation (OKFN). It assembles contributions of the workshop participants and, beyond this, it summarizes initial steps in the formation of a Linked Open Data cloud of linguistic resources, the Linguistic Linked Open Data cloud (LLOD).

Heidelberg and Leipzig,
December 2011

Christian Chiarcos
Sebastian Nordhoff
Sebastian Hellmann

Acknowledgements

We would like to thank the organizers of the DGfS for supporting our work by integrating the workshop on Linked Data in Linguistics as a part of the 34th Annual Meeting of the German Linguistic Society (Deutsche Gesellschaft für Sprachwissenschaft, DGfS), March 7th-9th, 2012, in Frankfurt a. M., Germany. Further, we thank the Max-Planck Institute for Evolutionary Anthropology Leipzig and the LOD2 project – a large-scale integrating project co-funded by the European Commission within the FP7 Information and Communication Technologies Work Programme (Grant Agreement No. 257943) – at the Agile Knowledge Engineering and Semantic Web (AKSW) Lab of the University of Leipzig for their financial and organizational support.

We would like to express our gratitude to the participants of the Workshop on Linked Data in Linguistics, whose contributions are assembled in this volume, for sharing their ideas, insights and/or resources. Due to organizational reasons, a considerable number of papers could not be accepted, and we would also like to thank the authors of these contributions, hoping that they stay in contact with us and keep on working towards the application of the Linked Data paradigm in their research.

The choice of papers was not always easy, and judging only on the quality or innovativeness of the submissions, we would have liked to include more of them. The selection process did, however, not only rely on these two criteria, but also involved other goals, partially conflicting with each other, such as the representativeness for the respective community, maturity of the approach, and the overall goal to illustrate the diversity of recent approaches. We thank the program committee (a list is included after this preface) for their invaluable support and engagement in this process and their feedback on the different contributions.

We also thank our invited speakers, Nancy Ide and Martin Haspelmath, for presenting their work and sharing their experience from two of the primary prospective fields of applications of the Linked Data paradigm in Linguistics, i.e., Natural Language Processing and typological research. Additionally, Christian Kreutz introduced the Open Knowledge Foundation, i.e., the host organization of the Open Linguistics Working Group (OWLG) that organized the workshop. We also thank the members of the OWLG who indirectly contributed to the workshop and to this

volume through discussions and the development of a common vision of a Linguistic Linked Open Data cloud.

Finally, we would like to thank Springer, for support and preparing the work.

Reviewing Committee

Anthony Aristar	Eastern Michigan University
Emily M. Bender	University of Washington
Hans-Jörg Bibiko	MPI-EVA Leipzig
Philipp Cimiano	CITEC, Universität Bielefeld
Alexis Dimitriadis	Universiteit Utrecht
Caroline Féry	Universität Frankfurt
Daniel Fleischhacker	Universität Mannheim
Jeff Good	University at Buffalo
Harald Hammarström	MPI-EVA Leipzig
Kees Hengeveld	Universiteit Amsterdam
Ernesto William de Luca	DAI-Lab, Technische Universität Berlin
Harald Lüngen	IDS Mannheim
Lutz Maicher	Fraunhofer MOEZ
John McCrae	CITEC, Universität Bielefeld
Gerard de Melo	MPI for Informatics, Saarbrücken
Pablo Mendes	FU Berlin
Steven Moran	University of Washington
Axel-C. Ngonga Ngomo	Universität Leipzig
Antonio Pareja-Lora	Universidad Complutense de Madrid
Cornelius Puschmann	Heinrich-Heine-Universität Düsseldorf
Felix Sasaki	DFKI Berlin, FH Potsdam
Stavros Skopeteas	Universität Bielefeld
Dennis Spohr	CITEC, Universität Bielefeld
Johanna Völker	Universität Mannheim
Menzo Windhouwer	MPI Nijmegen / Universiteit Amsterdam
Alena Witzlack-Makarevich	University of Zurich

Contents

Introduction and Overview 1
Christian Chiarcos, Sebastian Hellmann, and Sebastian Nordhoff

Part I Lexical Resources

Treating Dictionaries as a Linked-Data Corpus 15
Peter Bouda and Michael Cysouw

Integrating WordNet and Wiktionary with *lemon* 25
John McCrae, Elena Montiel-Ponsoda, and Philipp Cimiano

Integrating Lexical Resources Through an Aligned Lemma List 35
Axel Herold, Lothar Lemnitzer, and Alexander Geyken

Linking Localisation and Language Resources 45
David Lewis, Alexander O'Connor, Sebastien Molines, Leroy Finn, Dominic Jones, Stephen Curran, and Séamus Lawless

Part II Corpus Building and Annotation

Reusing Linguistic Resources: Tasks and Goals for a Linked Data Approach 57
Marieke van Erp

A Discourse Information Radio News Database for Linguistic Analysis . . 65
Kerstin Eckart, Arndt Riester, and Katrin Schweitzer

Integrating Treebank Annotation and User Activity in Translation Research 77
Michael Carl and Henrik Høeg Müller

Creating Linked Data for the Interdisciplinary International Collaborative Study of Language Acquisition and Use: Achievements and Challenges of a New Virtual Linguistics Lab 85
María Blume, Suzanne Flynn, and Barbara Lust

Part III Terminology Repositories and Knowledge Bases

Linking to Linguistic Data Categories in ISOcat 99
Menzo Windhouwer and Sue Ellen Wright

Towards Linked Language Data for Digital Humanities 109
Thierry Declerck, Piroska Lendvai, Karlheinz Mörth, Gerhard Budin, and Tamás Váradi

OntoLingAnnot's Ontologies: Facilitating Interoperable Linguistic Annotations (Up to the Pragmatic Level) 117
Antonio Pareja-Lora

Using Linked Data to Create a Typological Knowledge Base 129
Steven Moran

TYTO – A Collaborative Research Tool for Linked Linguistic Data 139
Andrea C. Schalley

Part IV Towards a Linguistic Linked Open Data Cloud: Recent Activities of the Open Linguistics Working Group

The Open Linguistics Working Group of the Open Knowledge Foundation 153
Christian Chiarcos, Sebastian Hellmann, and Sebastian Nordhoff

Interoperability of Corpora and Annotations 161
Christian Chiarcos

The German DBpedia: A Sense Repository for Linking Entities 181
Sebastian Hellmann, Claus Stadler, and Jens Lehmann

Linked Data for Linguistic Diversity Research: Glottolog/Langdoc and ASJP Online 191
Sebastian Nordhoff

Linking Linguistic Resources: Examples from the Open Linguistics Working Group 201
Christian Chiarcos, Sebastian Hellmann, and Sebastian Nordhoff

Acronyms

API	Application programming interface
ASJP	Automated Similarity Judgment Program
ATLAS	Architecture and Tools for Linguistic Analysis Systems
CES	Corpus Encoding Standard
CLARIN	Common Language Resources and Technology Infrastructure
CMDI	CLARIN MetaData Infrastructure
DCR	Data Category Registry, see ISOcat
DGfS	Deutsche Gesellschaft für Sprachwissenschaft ('German Linguistic Society')
HTTP	Hypertext Transfer Protocol
ISO	International Organization for Standardization
ISOcat	Data Category Registry maintained by ISO TC37/SC4
GOLD	General Ontology for Linguistic Description
GrAF	Graph Annotation Format, XML linearization of the LAF
LAF	Linguistic Annotation Framework, upcoming standard developed by ISO TC37/SC4
LDL	Workshop on Linked Data in Linguistics (LDL-2012)
lemon	LExicon Model for ONtologies
LMF	Lexical Markup Framework, standard for NLP lexicons and machine-readable lexicons developed by ISO TC37/SC4
LOD	Linked Open Data
LLOD	Linguistic Linked Open Data cloud
MT	Machine Translation
NLP	Natural Language Processing
OKFN	Open Knowledge Foundation
OLAC	Open Language Archive Community
OWL	Web Ontology Language, RDF extension for ontologies
OWL/DL	OWL dialect to represent Description Logics
OWLG	Open Linguistics Working Group
PAULA	Potsdamer Austauschformat für Linguistische Annotationen ('Potsdam exchange format for linguistic annotations')

PID	Persistent Identifier
RDF	Resource Description Framework
RDFa	Resource Description Framework in attributes, a collection of attributes and processing rules for extending XHTML to support RDF
RDFS	RDF Schema
SGML	Standard Generalized Markup Language, predecessor of XML
SKOS	Simple Knowledge Organization Scheme, RDF extension for knowledge representation
SMT	Statistical Machine Translation
SPARQL	SPARQL Protocol and RDF Query language
SQL	Structured Query Language
TBX	TermBase eXchange, an XML format designed to allow the exchange of terminology databases between tools, developed on the basis of TMX
TEI	Text Encoding Initiative
TMX	Translation Memory eXchange, an XML format for the exchange of translation memory data
URI	Uniform Resource Identifier
URL	Uniform Resource Locator
XLIFF	XML Localization Interchange Format
XML	eXtensible Markup Language