Semantic Web 1 (0) 1–39
IOS Press

Multilinguality and LLOD: A Survey Across Linguistic Description Levels

Dagmar Gromann [a,*], Elena-Simona Apostol [b,o], Christian Chiarcos [c], Marco Cremaschi [d], Jorge Gracia [e], Katerina Gkirtzou [f], Chaya Liebeskind [g], Verginica Mititelu [h], Liudmila Mockiene [i], Michael Rosner [j], Ineke Schuurman [k], Gilles Sérasset [l], Purificação Silvano [m], Blerina Spahiu [d], Ciprian-Octavian Truică [b,o], Andrius Utka [n] and Giedre Valunaite Oleskeviciene [i]

[a] *Centre for Translation Studies, University of Vienna, Austria*
E-mail: dagmar.gromann@gmail.com
[b] *Computer Science and Engineering Department, University Politehnica of Bucharest, Romania*
E-mails: elena.apostol@upb.ro, ciprian.truica@upb.ro
[c] *Institute for Digital Humanities, University of Cologne, Germany*
E-mail: christian.chiarcos@gmail.com
[d] *Dipartimento di Informatica Sistemistica e Comunicazione, Università degli Studi di Milano, Italy*
E-mails: cremarco@gmail.com, spahiu.blerina@gmail.com
[e] *Aragon Institute of Engineering Research, University of Zaragoza, Spain*
E-mail: jogracia@unizar.es
[f] *Institute for Language and Speech Processing, "Athena" Research Center, Greece*
E-mail: katerina.gkirtzou@athenarc.gr
[g] *Department of Computer Science, Jerusalem College of Technology, Israel*
E-mail: liebchaya@gmail.com
[h] *Research Institute of Artificial Intelligence, Romanian Academy, Romania*
E-mail: vergi@racai.ro
[i] *Institute of Humanities, Mykolas Romeris University, Lithuania*
E-mails: liudmila@mruni.eu, gvalunaite@mruni.eu
[j] *Department of Artificial Intelligence, University of Malta, Malta*
E-mail: mike.rosner@um.edu.mt
[k] *Formal and Computational Linguistics, KU Leuven, Belgium*
E-mail: ineke.schuurman@ccl.kuleuven.be
[l] *Laboratoire d'Informatique de Grenoble, Université Grenoble Alpes, France*
E-mail: gilles.serasset@imag.fr
[m] *Department of Portuguese and Romance Studies, University of Porto, Portugal*
E-mail: msilvano@letras.up.pt
[n] *Interdisciplinary Digital Resources Driven Research Institute, Vytautas Magnus University, Lithuania*
E-mail: andrius.utka@vdu.lt
[o] *Department of Information Technology, Uppsala University, Sweden*
E-mails: elena-simona.apostol@it.uu.se, ciprian-octavian.truica@it.uu.se

Editors: First Editor, University or Company name, Country; Second Editor, First University or Company name, Country and Second University or Company name, Country
Solicited reviews: First Solicited reviewer, University or Company name, Country; anonymous reviewer
Open review: First Open Reviewer, University or Company name, Country

*Corresponding author. E-mail: dagmar.gromann@gmail.com.

Abstract. Limited accessibility to language resources and technologies challenges communities of speakers of any language other than English. Linguistic Linked (Open) Data (LLOD) holds the promise to ease the creation, linking, and reuse of multilingual linguistic data across distributed and heterogeneous resources. However, individual language resources and technologies accommodate or target different linguistic description levels, e.g. morphology, syntax, phonology, and pragmatics. In this comprehensive survey, the state-of-the-art of multilinguality and LLOD is being represented with a particular focus on linguistic description levels, identifying open challenges and gaps as well as proposing an ideal ecosystem for multilingual LLOD across description levels. This survey seeks to contribute an introductory text for newcomers to the field of multilingual LLOD, uncover gaps and challenges to be tackled by the LLOD community in reference to linguistic description levels, and present a solid basis for a future best practice of multilingual LLOD across description levels.

Keywords: Multilinguality, Linguistic Linked Data, Linguistic Description Levels, Systematic Survey

1. Introduction

Human languages are manifold, they shape communities and their interaction with each other, with national institutions or with the global economy. They also conceptualise the world, since categories and patterns of use of any particular language have an impact on its speakers [1]. Language pluralism is thus an integral part of our universal cultural heritage and a defining aspect of social and political structures. Furthermore, digital language data capture and document language use in a community at a specific moment in time, hence representing important cultural assets [2]. At the same time, the interaction of language communities with the globalised world puts speaker communities under pressure from major languages. Budin and Melby [3] identified legal, economic, information, technical, and methodological barriers to the interoperability of language resources. A limited accessibility of language resources and technologies represents a challenge for speakers, and with the rise of digital mass communication, the Internet, and wide-spread use of online services, the language barrier becomes a pressing issue for speakers of any language other than English. In particular for low-resource languages, the consolidation of existing data and the development of technologies to facilitate information integration from different multilingual resources are thus essential first steps for exploiting possible synergies and better services on this basis.

High-quality digital language data and resources are vital to a variety of research areas, such as linguistics, the study of low-resource languages, and language typologies. Such data are equally important for a number of downstream applications from Natural Language Processing (NLP) to learning structured knowledge from text. The creation, linking, and reuse of multilingual linguistic data is complex due to differences in theoretical underpinnings, representation formats, and annotation and metadata coverage. In particular, differences in linguistic description levels need to be considered, such as the morphological, syntactic, lexical, and other (see Section 4), i.e, in the form of a technology that is sufficiently generic to be applied to all levels of linguistic description and capable of integrating information from different data providers, e.g., from national research infrastructures used for hosting their respective language resources.

With this objective in mind, Chiarcos et al. [4] introduced the notion of Linguistic Linked (Open) Data (LLOD)[1] for applications in the context of language technology and multilinguality challenges, that is, to use the Linked Open Data (LOD) [5] ecosystem, technologies and formalisms to establish interoperability between language resources and to integrate information from various, distributed and heterogeneous resources. In particular, publishing linguistic data in this way allows resources and their components to be globally and uniquely identified such that they can be retrieved through standard Web protocols. Moreover, resources can be easily linked to one another in a uniform fashion, and the development and application of commonly shared, open vocabularies is strongly encouraged in this community, so that resources become structurally and conceptually interoperable, re-usable and sustainable,

[1] "Open" is in brackets since proprietary data can also be published as linked data. We use LLOD to refer to the technology and the use of open, community-maintained vocabularies, regardless of the licensing and availability of the resources this is applied to.

and – particularly important for multilingual applications – this facilitates the creation and querying of links across resources from different languages, across different levels of description or by different providers [6].

This article represents a comprehensive survey of the state-of-the-art in multilinguality and LLOD with a particular focus on support for different linguistic description levels in order to identify open challenges and gaps. Bosque-Gil et al. [7] and more recently Khan et al. [8] present surveys on modeling linguistic data as LLOD, where the former identify phonetics and phonology as well as dialogue structures as still under-represented. In this more comprehensive and recent survey we can confirm these findings and additionally identify pragmatics as a level with rather low coverage to date. To the best of our knowledge this is the first systematic survey of existing research and practices of linguistic description levels in multilingual LLOD resources. Building on the Preferred Reporting Items for Systematic reviews and Meta-Analyses (PRISMA) [9] method to conduct and report systematic reviews and a team of 15 experts in linguistics and LLOD, this article aims to:

- provide guidance for researchers and practitioners on available approaches for supporting specific linguistic description levels in the LLOD;
- identify open challenges and gaps in the support of linguistic description levels across multilingual LLOD resources; and to
- present a solid basis for a future best practice on how to represent, model, and link different linguistic description levels across multilingual LLOD resources.

The article is structured as follows: Section 2 introduces the preliminaries of multilinguality and LLOD. Section 3 then describes the methodology and statistical results of the conducted survey. Sections 4 and 5 detail the findings from our survey, where the former focuses on models and types of linguistic description levels covered, while the latter concerns types of language resources with their linguistic description levels and their use. Section 6 unites challenges that were identified based on this survey with challenges that derive from the experience of the group of experts authoring this article. Finally, prior to concluding remarks, Section 7 proposes an ideal ecosystem for multilingual LLOD, addressing general challenges that need to be addressed by the (L)LOD community as well as particular challenges that pertain to multilinguality and LLOD.

2. Background and Motivation: Multilinguality and LLOD

The two concepts of linking and multilinguality are of fundamental importance because they relate strongly to the distribution of data according to FAIR[2] principles and in particular to interoperability between datasets, which is one of the key benefits claimed for the use of LLOD. Linking clearly allows data silos to be connected together to promote interoperability at different levels of granularity. It also offers a way to lift any barriers imposed by the language-specific nature of data. It is no surprise that this fundamental aspect of multilinguality dearly appealed to researchers in semantics and language who saw it as an opportunity to overcome the "monolingual islands" effect [11, 12], i.e., the problem of connecting and accessing data expressed in different languages. Below we further examine the concepts of multilinguality and LLOD.

2.1. Linking Data to Language

In the context of web technologies, the most widely adopted solution to the issue of how to perform this linking is the application of the Resource Description Framework (RDF) [13] and Linked Data [14]. Cimiano et al. [15] present the semantics of the RDF model, which was created in late 1990s, to represent linked data and knowledge in a machine-readable manner, and its most common formats for serialisation, N-Triples, Turtle, XML and JSON-LD, which enable publishing RDF data on the Web. The authors also give an overview of the Web Ontology Language (OWL) and SPARQL, the standard language for querying RDF data. With the development of commonly used vocabularies for language resources, especially for the lexical domain (OntoLex-Lemon [16, 17]), the so called LLOD cloud has been developed [4, 18] as an aggregator of language resources available as LOD, and, subsequently,

[2]FAIR data principles are intended for improving Findability, Accessibility, Interoperability and Reusability [10].

great potential has been recognised in the use of this technology to establish interoperability between existing resources for language technology, especially in applications that have previously been tackled by means of graph technologies or feature structures, such as lexical data or linguistic annotation [19–21]. Also, the SKOS standard for representing structured controlled vocabulary is widely used for the representation of multilingual LLOD [2, 22] and SKOS-XL[3] is used for representing links across multilingual resources [23]. LLOD results from the convergence of three long-standing trends in software development and language technology, i.e., open data, linked data and language resource interoperability. The LLOD cloud emerged from the growing number of linguistic resources independently published in accordance with LOD principles, and from the will to link them across languages [12], with benefits in the areas of representation and modelling, structural interoperability, conceptual interoperability, federation, dynamicity, and ecosystem [24, 25]. LLOD is an exemplary application of FAIRness in science [18], so that after the proposal of the FAIR Guiding Principles for scientific data management and stewardship [10], this trend intensified even further.

Multilinguality has always been a central aspect of LLOD development. Initially, most LOD resources adopted *language agnostic ontologies* that were associated with language data only by means of `rdfs:label`, a property designed to provide a human-readable version of a resource name. In this context, the main problem was to identify language, dialect, or variants of such labels. This was quickly followed by other problems associated with linguistic characteristics of labels – how to access the respective lexical entry, related word senses, etc. For these purposes, the simple use of `rdfs:label` was abandoned in favour of a structured, reified representation of natural language labels, thus permitting sufficiently detailed descriptions of their linguistic behaviour to be expressed using data models such as SKOS-XL, or OntoLex, elaborate domain vocabularies such as GOLD [26], LexInfo [27] and OLiA [28]. Together, these form a commonly accepted framework to accommodate aspects of multilinguality, and the transition from simple labels to structured linguistic descriptions is the hallmark of the establishment of LLOD as a separate branch of LOD technologies.

With the increasing number of available multilingual language resources as LLOD, the question of adequate support not only for multiple languages but different description levels in individual resources becomes more and more pressing. Several approaches exist for tracking information about the same item across different data sources exploiting links, such as `owl:sameAs` [29–31], providing multilingual access to information in ontologies [31] or multilingual contexts to cultural heritage objects [32], and enabling multilingual querying over multilingual knowledge graphs [33]. Furthermore, several works [34–37] have highlighted that LLOD can pave the way for better discovery and connectivity of linguistic data of under-resourced languages, and for new ways to preserve cultural diversity.

As a result of these trends we find ourselves today in a situation where the semantic layer is no longer the only bridge between languages. Translations are, in principle, possible via the linguistic layer either statically, through pre-computed cross-lingual links, or dynamically, by computing such links on the fly. Furthermore, because the computation of such translations can exploit a wide range of linguistic resources available in the cloud, they can be sensitive to linguistic and cultural context and can exhibit a degree of finesse and nuance not realisable from a purely semantic perspective.

The full potential of this approach is yet to be fully determined, which is why we feel it is opportune to carry out a systematic survey which has to take into account the complex interplay of progress between (i) the different levels of linguistic description that make up the layer of linguistic information present in the LLOD (ii) the representations and models that are used to express these different levels and (iii) the use cases in which these have been realised.

2.2. The Concept of Multilinguality

The notion of multilinguality is pervasive throughout the LLOD literature, and its meaning is generally taken for granted. However, close examination of the way the concept is used reveals a variety of accepted meanings. Some idea of this variety can be revealed by observing that the things that are frequently cited as being "multilingual" fall broadly into three categories: (i) language resources, (ii) tools and services, and (iii) knowledge-based structures, i.e., ontologies, knowledge graphs, taxonomies and databases.

[3] http://www.w3.org/TR/skos-reference/skos-xl.html

Language Resources are characterised by static linguistic content, i.e., content that, being linguistic, belongs to a given natural language. Prototypical examples of such resources are text corpora, wordlists and lexicons. All of these resources structure natural language at a higher level (e.g. paragraphs, sentences) and at a lower level (e.g. characters or sound patterns). This structure may be expressed in the form of explicit annotations.

Services and tools are not static but display behaviours having inputs and outputs. So, for example, a tagging service takes a textual input and outputs annotations that include part-of-speech (POS) information. A Named Entity Recognition (NER) service does the same but with named entities. With some services we are more concerned with the behaviour itself than with the input/output relations. So in the case of a chatbot there is the tendency to focus on the quality and feel of the user experience rather than on the overall input/output relation. However, even in such a case there still has to be both input and output and furthermore it is language-dependent.

Knowledge-based structures comprise, on the one hand, descriptions at conceptual level (systems of concepts and relations between concepts) and, on the other, instances of those concepts. Such structures are not language resources in the classical sense because the concepts and their instances are not natural language words. However, to aid understanding, they are often given names which are natural language words, and this may lead to the interpretation that they represent linguistic data similar to language resources.

2.3. What Makes the LLOD Cloud Multilingual

It is a truism that what distinguishes LLOD from mere LOD is that the data has the essential character of *being linguistically relevant*. In the context of LLOD, "a dataset is linguistically relevant if it provides or describes language data that can be used for the purpose of linguistic research or natural language processing" [2, p. 33]. In trying to pin down what makes LLOD multilingual, we need to clarify what that connection might be. As hinted above, there may be several possible kinds of connection depending on the nature of the multilingual entity: resource; tool; knowledge structure. Next, monolinguality is discussed for each of these cases. This is expanded to cover multilinguality.

2.3.1. Multilingual Resources

A resource is monolingual if its contents relate to one language. For example, a corpus of Italian text or an Italian wordlist is monolingual because it contains words which belong to the Italian language. This can be generalised: a resource is multilingual if it relates to two or more languages. A prototypical example would be a code-switching corpus, e.g. [38] whose words derive from both English and Mandarin. A resource can also be multilingual if it is composed of several monolingual subparts belonging to different languages. This is consistent with Schmidt and Wörner [39], for whom a multilingual resource is "any systematic collection of empirical language data enabling linguists to carry out analyses of multilingual individuals, multilingual societies or multilingual communication".

Essentially, the LLOD cloud is multilingual because it includes corpora and other resources that contain data in a variety of different languages. A separate issue, and one of great importance, is how that information is actually represented. Ultimately, it has to bottom out in the association of an entity of some kind (e.g. a string) with a universally accepted language label.

2.3.2. Multilingual Services and Tools

A monolingual service or tool is characterised by three things: input, outputs and behaviours. In many ways inputs and outputs resemble mini language-corpora in that they bottom out in natural language strings. Accordingly, a service or tool will be deemed monolingual if it operates over inputs and outputs that (like monolingual corpora) are both associated with the same unique natural language. Expanding this to the multilingual case, there are several possibilities: (i) input and output are in different languages (e.g a translation service); (ii) same service can be applied to input/output in same language but for different languages (e.g. EN-EN and FR-FR summarisation); (iii) various combinations of (i) and (ii). It is also possible to envisage NLP services where either input or output is not in natural language as such but in some other form, such as a parse tree or an abstract meaning representation. The linguality of such structures are discussed in the next section.

2.3.3. Multilingual Knowledge Structure

Examples of knowledge structures are ontologies, propositions, taxonomies, etc. Items in this class have several distinguishing characteristics. First, they can be represented directly using LLOD machinery (e.g. using RDF, shared vocabulary, naming with URIs, links to other resources). Second, they are primarily *conceptual*, not linguistic - i.e. they concern concepts and instances of concepts rather than language strings. A taxonomy, for example, is a classification scheme whose elements are connected by relations such as "IsA" and "hypernym". Third, despite being conceptual, they generally connect to language in some way for the sake of understandability. However that connection is less direct than for a string. Thus we can refer to the concept of a dog using the English string "dog" so that every English speaker will understand what we are referring to. Knowledge structures are thus at least monolingual. Clearly the example can be generalised to include strings in as many other languages as we like, and it is in this sense that we understand what it is for a knowledge structure to be multilingual.

2.4. Multilinguality vs. Language Independence

Many linguistic or lexical approaches have claimed to be multilingual, because they are not rooted in a specific language. It would be more precise to say that they are *language independent*. LOD extends this idea because not only are its design principles language independent, but they emphasise in addition two key points: (i) *reuse* of existing conceptual vocabularies, rather than creation of additional versions of the same concepts for different languages, and (ii) *extension* of existing vocabularies when they do not exactly fit the author's need, together with a semantic description of such an extension and its motivation.

When these points are realised, even purely monolingual datasets (e.g. monolingual annotated corpora) can reuse the same set of linguistic features as other datasets in other languages. Hence, two independent monolingual corpora may be queried for common patterns, using a common vocabulary, leading to a multilingual use case or service based on monolingual data. In this way, LOD enables multilinguality as interoperability between languages, even on resources or services that are initially designed as monolingual.

Elaborating a little bit further on the example of a monolingual annotated corpus, we should also stress that, even if no common vocabulary is fine grained enough for the representation of some peculiarities of the represented language, it is still possible for the author to further refine existing data categories and achieve linguistic felicity in the language description while still allowing interoperability with other language resources or services. We note that in the domain of morpho-syntactic annotation, Universal Dependencies [40] strive to achieve something similar: cross-linguistic consistency of annotation, while still permitting language-specific extensions when necessary.

Applicability to different languages leads to real multilinguality, due to the design principles of LLOD. With multilinguality achieved through interoperability of languages, LOD is able to express the best of both worlds: very fine description of a specific language (linguistic felicity) + linking through shared vocabularies by way of refinement/extensions of existing shared vocabularies. Before discussing approaches to create, represent, and reuse multilingual language data building on LLOD principles, we first introduce our approach to implementing this systematic review.

3. Approach of Systematic Review

This section gives a detailed description of the methodology we applied to our systematic literature review, based on the well established PRISMA method [9], and details on the obtained results of the systematic review that serve as a basis for the comprehensive analysis in the following sections.

3.1. Methodology

The objective of this systematic review is to provide a synthesis on the state of knowledge (Sections 4 and 5) and suggestions for priorities of future research (Section 6 and 7). The PRISMA method has specifically been designed to provide detailed reporting guidelines for such reviews to ensure a comparable and comprehensive result. This method generally consists of three stages:

- Identification
- Screening
- Inclusion

3.1.1. Identification

In order to optimise our search in publication databases, a set of keywords was jointly defined by a group of, in total, 15 experts. Each keyword represented a composition of multilingual, multilinguality, multilingualism or cross-linguistic, cross-lingual and prototypical search terms for LOD, e.g. RDF, linked data, web or simply "multilingual data". In addition, we explicitly included linguistic description levels in the keywords, i.e., pragmatics, syntax, semantics, lexical, discourse analysis, phonology, phonetics, and morphology. In total, 41 individual, e.g. ["multilingual LLOD"], and compositions of keywords, e.g. ["multilingual data" AND "representation"], were jointly identified as relevant. The keywords were collected in a document and discussed in several meetings as well as initially submitted to one search platform to test their potential return, i.e., if there was no result the keyword was excluded from further steps. In a second step, the keywords were rated on a scale from 1 to 10 by 6 experts, where 1 signified not relevant and 10 denoted highly relevant for this search. We calculated an average for each keyword/keyword combination from these scores to obtain a final relevance score.

These keywords represented a starting point for an extensive search on several publication platforms, which the same group of experts jointly identified as important to this task. The following search platforms for scientific publications were utilised in the proposed approach:

- Scopus
- Web of Science
- DBLP
- Google Scholar

The time period was set from 2009 until 2021 for this search, which focuses our survey on more recent works. We additionally assumed that important publications before 2009 would be included in review papers that fall within the time period we selected. To reduce the number of resulting publications to a manageable number of papers to be read by the 15 experts of this research endeavour, each paper was ranked by times of occurrences across platforms and keyword ranking building on the expert scores introduced above. The final score for each paper was calculated by taking the score for each search keyword the paper resulted from and multiplying it with the times of occurrences across platforms, finally summing the individual multiplied keyword scores. For instance, Paper No. 1 was found with the keyword ["multilingual LLOD"] with an expert score of 9.17 three times across platforms resulting in a score of 27.51. The same paper also resulted from the keyword ["multilingual information"] with an average expert score of 4.17 one time, which makes the total score for this paper 31.68 in the final ranking. This approach clearly favours papers resulting from several keywords that were ranked with a high expert score.

The extensive search was supplemented with snowballing, i.e., exploration for more recent publications citing central works we identified within our result corpus. In parallel, a reference repository of publications that this group of experts considered central to this topic was compiled. This reference repository serves as a gold standard to validate our semi-automated keyword-based search strategy. We have evaluated to which degree the result corpus of the latter contains publications from the reference repository.

3.1.2. Screening

The top-rated papers from the Identification step were manually annotated each by two experts. A crucial and central qualifying question for the screening process was which linguistic description levels are addressed/described in each publication. Furthermore, the criteria for this Screening step were the relevance of the publication to the topic of multilingual linguistic linked data and its thematic categorisation by representation, approach or standardisation. If one or two annotators marked a paper as "unsure", i.e., not clearly central to this survey but probably to be considered, a third expert decided on the publication's relevance.

To distribute the final set that resulted from this initial screening among experts, we performed an annotation process with pre-defined categories based on their title, abstract and keywords. Only if the categorisation based on these three components of publications was not possible, the full text had to be consulted at this stage. The categories

Table 1
Tags for expert annotation of result set

Type	Categories	Examples
Generic tags	Application	
	Representation	
	Resource	
	Use case	
Specific tags	linguistic description levels	phonology, lexical level, syntax, semantics, pragmatics, terminology, discourse analysis, co-reference
	approach	e.g. bilingual linking
	standard/format	e.g. OntoLex, OWL, SKOS, RDF, TEI, LMF, TBX, UMLS, etc. or "several" if not one specific

Table 2
Types and numbers of clusters with number of publications per cluster and experts

Label	No. Publications	No. Experts
application	15	2
BabelNet	5	1
literature reviews	5	1
LLOD infrastructure	4	1
morphology	5	1
OntoLex-Lemon	25	3
overview publications	6	1
representation	12	2
resources	12	2
standards	5	1
under-resourced languages	4	1
use cases	12	2
Total	110	18

for this final step were divided into generic and specific annotation tags represented in Table 1, where the specific tag of linguistic description level had to be assigned to all publications.

For generic tags, the category was only assigned if relevant for a given publication. For specific tags, each of the three categories and a respective value exemplified in Table 1 was assigned. This annotation with generic and specific tags provided the basis for clustering the result set, assigning each cluster a specific label. The clusters served the purpose to decide on the relevance of an individual publication by comparison to other publications on the same topic, perform targeted snowballing and ensure that experts can search for more recent publications on the specific topic, mitigating the risk to miss important contributions. Furthermore, it facilitated the distribution of the workload among the experts.

To decide on the eligibility of publications, each cluster was assigned to one, two or three of the experts of this work, depending on the size of the cluster. A cluster in our case is a grouping of papers based on their identical or similar tags. Very large clusters would be assigned to three experts, very small clusters to only one expert. Some clusters that contained a considerable number of papers on a specific subtopic, e.g. OntoLex-Lemon, were further subdivided. Table 2 shows the types of labels and number of clusters, the number of papers contained in each cluster and the number of experts that worked on each cluster. As you can see in Table 2 some of the 15 experts were assigned to more than one cluster.

3.1.3. Inclusion

This section describes our methods for identifying the final subset of publications to be included in this review. The first and foremost criteria for inclusion were that publications are:

- directly related to multilingual linked data
- published in English
- peer-reviewed (guaranteed by the publication venue)

The explicit decision which publications to report was taken by the experts of the individual clusters, where specific papers would be discussed with other experts if the decision was not clear. Snowballing, that is, checking citations in our result set on important works, and complementing the result set with additional more recent publications, further increased the number of publications considered for this survey.

Inclusion was designed as a two-step process. In the first step, experts assigned to a specific topic, i.e., a cluster in our case, prepared a written summary of topic-specific publications, dividing the contents into the topics that now represent Sections 4 to 5 of this article for uniformity. In the second step, the individual sections of each cluster summary was synthesised into the sections of this article.

3.2. Results

The total number of papers for each stage of the survey methodology is represented in Fig. 1. In the Identification stage, we identified 41 keywords that were ranked by 6 experts according to their relevance. The Spearman correlation for this ranking step was 0.632 across all six expert rankings, thus providing a strong correlation. The keyword scores provided the basis for ranking the papers, adding up scores of a paper depending on from which keyword it was returned. In total from 41 keywords a list of 25,074 papers were returned.

Given the number of people involved and the time available to annotate papers, we had to limit the result set to annotate. To this end, after removing duplicates, the result set was ranked by keyword-based score and the top-ranked publications were inspected to determine a cutoff score. This cutoff turned out to be a score of 37, after which publications started to get less relevant to our topic, limiting the result set to be screened to 210 publications. For comparison, the top-ranked publication obtained a ranking score of 155.19. Manually screening and annotating this reduced result set further decreased the number to 110 publications after the screening phase (see Section 3.1.2), removing not directly relevant or duplicate publications. This manual annotation first involved assessing whether a paper is relevant (1), not relevant (0) or the annotator was unsure about its relevance (2). The inter-rater reliability score for this rating resulted in a moderate kappa value of 0.495, mostly due to the fact that many times one rater was sure about relevance, while the second annotator was unsure, providing a 2. In cases were a 2 was assigned, a third annotator would determine whether to include the publication or not. This detailed screening stage led to the exclusion of 14 more papers, 4 of which were superseded by newer publications by the same authors, 6 were closely related to other use cases, e.g., on BabelNet or OntoLex-lemon, and 4 were finally deemed not closely related to linguistic description levels.

The size of the clusters varied between 4 and 25 publications, the smallest was related to the tag LLOD infrastructure, the largest to the specific representation format and standard *OntoLex* and its predecessor *Lemon* [16] as represented in Table 2. Summaries of these clusters were prepared by experts and structured by the topics and sections in this article. Not all of these topics would be covered by each of the clusters, e.g. the topic of morphology did not explicitly address other linguistic description levels.

In terms of gold standard comparison, from the 10 papers manually selected as highly relevant by experts, only 6 were included in our final result set. This confirms our intuition that this method should be extended by performing snowballing and further investigation on the individual linguistic description levels, which we performed when deemed necessary. The final number of papers included in this survey comprises 203 publications. We kept references to individual book chapters of a monograph if these were part of our result set and referenced them accordingly in this work.

All publications surveyed and added by means of snowballing and exploring more recent publications are finally discussed in the following Sections 4 and 5. First, we present approaches specific to individual linguistic description levels. Second, resources, their uses and representation models are discussed. In Section 6 and 7, we draw conclud-

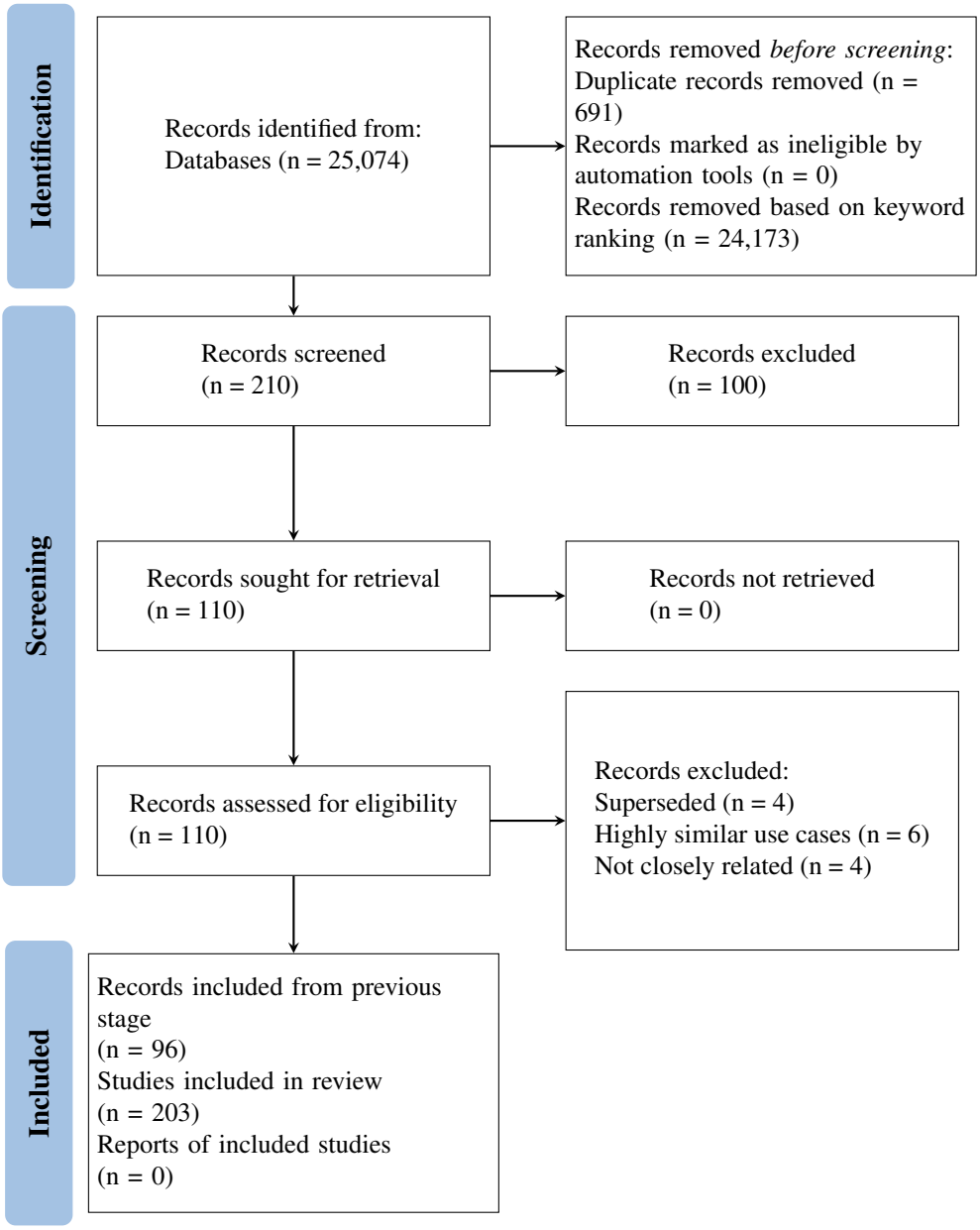

Fig. 1. PRISMA 2020 Flow Diagram

ing challenges from the survey analysis as well as our own professional experiences and discuss a potential ideal ecosystem for LLOD with respect to multilingual data and linguistic description levels.

4. Linguistic Description Levels: State-of-the-Art

In this section, we analyse the results of our literature analysis along the following linguistic description levels:

- Lexical Semantics
- Syntax and Morphology

- Pragmatics
- Lexicography
- Phonetics and Phonology
- Translation and Terminology
- Etymology and Diachronicity

One recurring and predominant model for representing linguistic information as linked data at different linguistic description levels is OntoLex-Lemon. Thus, several of the approaches covered in this section represent extensions of OntoLex-Lemon (see [41, 42] for an overview on such extensions). It also occupies a central role as representation mechanism in the integration of resources and services into complex language technology-processing pipelines [43]. Nevertheless, the objective of this section is to provide a general overview of approaches to describe different linguistic description levels within the context of multilingual linked data. This overview serves the purpose to see which levels have been well covered in the literature and which ones might require more attention as well as to identify open challenges.

It should be noted that the majority of reviewed papers do not refer to specific linguistic descriptive levels, but rather have generic references to "linguistic data", "lexical data", "language annotations", "annotated corpora", etc. Such generic references typically include several linguistic description levels that deal with written language, e.g. morphology, syntax, (lexical) semantics, etc. Bosque-Gil et al. [7] explicitly touch upon representation of specific linguistic levels, i.e., phonetics and phonology, morphology, syntax, semantics, semiotics, discourse, and specific branches of linguistics, i.e., historical linguistics, lexicography, typology and cross-linguistic studies, terminology. Bosque-Gil et al. [7] observe that "phonetics and phonology remain two areas with relatively low coverage in the LLOD cloud" as well as dialogue structure. Our more comprehensive and more recent survey can confirm this finding based on the coverage of description levels and number of papers in the result set on these description levels. Additionally, we identified a low coverage for pragmatics. While we touch upon modeling of linguistic data and different linguistic description levels in this and the following section, please consult Khan et al. [8] for a very comprehensive survey on the current state-of-the-art on modelling LLOD.

4.1. Lexical Semantics

Lexical semantics is the study of word meaning. Within the context of this article, we are interested in how word meaning in all its facets can be represented in LLOD. Several models to represent lexical data on the web have been defined:

- LingInfo [44]
- LexOnto [17]
- Linguistic Watermark framework [45, 46]
- Linguistic Information Repository (LIR) [47]
- Lexicon Model for Ontologies (*lemon*) [16] and its most recent version Ontolex-Lemon[4] [41]

These models made it possible to link the semantic information described in existing ontologies with the linguistic information necessary to link ontological concepts with their mentions in natural language data.

From these models, the OntoLex-Lemon predominantly surfaced in our result set, also in its preceding version *lemon* (see Table 2), including numerous applications and use cases (see Section 5). It was developed from the Lexicon Model for Ontologies (*lemon*), which builds on LIR [47], LexInfo[5] [27], the Lexical Markup Framework [48] and SKOS, and relies on standardisation efforts such as ISOcat metadata registry [49] and OLiA [28].

In the core model of OntoLex-Lemon, headwords are represented as lexical entry (`ontolex:LexicalEntry`), which can be either (single) words, multiword expressions or affixes (such as *un-*) [42]. The base linguistic form of the entry or lemma is called the canonical form. In case of multiword expressions, the decomposition module can be utilised to describe its internal structure and components. To represent the meaning of a lexical entry, it is linked to

[4] https://www.w3.org/2016/05/ontolex/
[5] https://www.lexinfo.net/

a lexical sense (`ontolex:LexicalSense`). This not only allows to represent different senses in connection to a single entry, but also to add additional information to the sense level, such as the status of use of a specific sense, e.g. outdated. Originally, OntoLex-Lemon was designed to represent lexical semantics in relation to ontologies, which is why lexical senses can `ontolex:reference` an element in an ontology. Alternatively, a conceptual model can be included within the lexicon. For instance, the OntoLex-Lemon representation of WordNet relies on synsets for a conceptual model [50]. One extension of lexical representation in OntoLex-Lemon on the lexical semantic layer is proposed in the form of Lexical Function Ontology Model (Lexfom) [51], which represents lexical functions as paradigmatic, e.g. antonymy, synonymy, meronymy, and syntagmatic, e.g. objective or subjective qualifications, relations between lexical units and senses.

The original lemon model [16] advanced in the context of the W3C OntoLex community group[6], resulting in the new OntoLex-Lemon model, published as a W3C report[7]. The W3C OntoLex community group remains an active one that further develops the Ontolex-Lemon model in order to extend its applicability. The group has recently aimed to develop four new modules [41] for Morphology (see also Section 4.2), Lexicography (see also Section 4.4), Etymology and Diachronicity (see also section 4.5) and lexico-syntactic categories. Most of the works on LLOD for under-resourced languages describe lexical data on the basis of the *lemon* model (either its initial version or the more recent OntoLex one). Additionally, other modules for extending OntoLex-Lemon have been proposed to address different types of linguistic information. For instance, Onyx [52] represents an extension of lemon to model emotion information and the emotion analysis process itself, which can also accommodate multilingual information.

A model for describing lexical semantics preceding and extended by OntoLex-Lemon is the Simple Knowledge Organisation System (SKOS) [53]. It is an RDF vocabulary designed to represent concept schemes and provide lexical information for thesauri and other types of controlled vocabularies. Lexical meaning is represented as `skos:Concept` that only requires a URI and an RDF type declaration. Lexical manifestations are added by means of three types of labels: preferred, alternative, and hidden. The last type serves to include obsolete or other forms for machine processing and searching that should not be visible or used otherwise. Concepts can then be organised hierarchically with broader/narrower relations and non-hierarchically with an associative relation. Directly attaching lexical strings to a concept fails to allow for separate metadata descriptions of the lexical semantic/conceptual and word level, which is a problem that was solved by introducing SKOS-XL, which separated these two levels. While SKOS publications were not directly part of our result set, publications utilising SKOS as a data model were included (see Section 5).

One alternative approach to represent lexical semantics in our result set is Framester [54], a data hub focused on broadening the FrameNet coverage of linguistic information and formal homogeneous linking of lexical and factual resources. Building on Fillmore's frame semantics [55] and Linguistic Linked Data principles, it acts as a hub between FrameNet, WordNet, VerbNet, BabelNet, DBpedia, DOLCE-Zero, and many other resources. It provides a two-layered (intensional-extensional) semantics for frames, semantic roles, semantic types, selectional restrictions, and other elements of lexical resources in OWL2. Any word or multiword can then evoke a frame, which can be a FrameNet frame or any other type of frame, such as a WordNet synset frame. While this approach allows for easy access via a SPARQL endpoint and a different representation model for lexical semantics, multilinguality is not explicitly considered and only covered in as far as the interlinked resources are multilingual.

From the perspective of linked data, all approaches to represent lexical semantic information agree that the conceptual or meaning level should be kept separate from the string or word level. This is important since additional information might only apply to one of these levels, e.g. part-of-speech relates rather to the lexical representation than to the meaning of a word. A separation of meaning and form is particularly important for representing multilingual information, such as equivalent words or multiwords across languages that represent the same meaning but require different metadata descriptions. While this section might not cover all proposed approaches to represent lexical semantics as LLD, it clearly shows the level of sophistication of existing approaches and that this linguistic description level might be one of the most developed in LLD.

[6] https://www.w3.org/community/ontolex
[7] https://www.w3.org/2016/05/ontolex/

4.2. Syntax and Morphology

Syntax guides the composition of words and morphemes into larger units of phrases and sentences. Morphology studies the composition of words, where inflectional morphology is concerned with affixes that carry grammatical meaning to fit words within specific grammatical contexts and derivational morphology relates to the formation of new words with changes to part-of-speeches and lexical meaning. One common way to represent syntactic and morphological information in relation to textual data and corpora is by means of annotation metadata. A very comprehensive ontology to formalise linguistic information in a machine-readable ontology for 75 language varieties is provided by the Ontologies of Linguistic Annotation (OLiA) [28], which covers morphology, morphosyntax, phrase structure syntax, and dependency syntax. Recently, OLiA has been utilised in Annohub [56], a method to harvest existing annotation schemes to provide an RDF-based platform for linguistic research.

In OntoLex-Lemon, the syntactic behaviour of headwords in the lexicon, i.e., lexical entries, can be described by means of syntactic frames and the number and type of arguments a lexical entry requires [42]. For instance, verbs that follow a transitive frame require a syntactic subject and a direct object. Morphemes can be represented as different forms of a lexical entry, e.g. singular and plural forms. A very specific scenario for re-using OntoLex-Lemon to model morphological and syntactic information is provided by [57], who target to represent annotations generated from language-learning content. As examples, the authors model a Spanish conjugation and an English syntax exercise as LLD.

One phenomenon at the syntax-semantics interface that we decided to include in this section for the purpose of this overview is that of *coreference*, which represents a binding phenomenon of elements within and across sentence, such as anaphora or coreferring noun phrases. Bryl et al. [58] explore the extraction of different surface forms from Wikipedia in order to enhance DBPedia entities with additional filtering steps, since these forms are important for disambiguation and coreference resolution. The additional filtering relies on string patterns and information from Wikidata and TF-IDF calculations. Prokofyev et al. [59] propose SANAPHOR, a system that identifies text mentions, which can be either entities, pronouns or determiners, and types them with a knowledge graph, such as DBpedia, in order to improve coreference clustering. In an extended SANAPHOR++ version [60], the authors extend the initial system to better handle ambiguous entities, e.g. *Paris* the city and person, novel entities, and integrate additional semantic features on the mentions.

Morphology still remains an under-explored aspect of LLOD. With the systematic review, we identified papers that address morphology in lexical resources [61–63], corpora [64–66] and in grammars [67] and as general modelling challenges [61, 68].

In all of these areas, a number of more recent publications have appeared, which we added after the systematic review. OntoLex-Lemon extensions for morphology initially focused on inflectional morphology and composition with limited support for derivational morphology. The Multilingual Morpheme Ontology (Mmoon) [61, 69] has been designed in a bottom-up approach to provide an exhaustive vocabulary for morphological inventories, partly inspired by current standards, tools and resources as applied in language documentation and linguistic typology. Its feature inventory incorporates a large number of terminological resources that are of considerable size in their own right (ISOcat, OLiA, LexInfo), which is why it has grown into a relatively large vocabulary. Mmoon [68] focuses on decomposition of entries and related word forms as well as morphological patterns that are used to form lexical entries and word forms. To this end, an extension of OntoLex-Lemon by 11 classes and properties has been proposed, the most central ones being `morph:Morph`, which has six subclasses of specific subtypes, such as AffixMorph and TransfixMorph, and a class `morph:MorphologicalPattern`, which describes the morphological building pattern of the entry and its related word forms.

Several future, additional features that should be addressed, such as ordering morphs, which is not strongly supported by the current RDF format, are discussed. Preliminary work in this sense is reported in Declerck et al. [70] which shows how the lexical representation and linking features of OntoLex-Lemon can be used to model morphological and ordering restrictions over the components of Multiword Expressions (MWEs), illustrated by examples from OdeNet, a German resource for lexical semantics. Because of the complexity of the vocabulary, it is lacking wide application, but it has been driving the development of the OntoLex-Morph module [68]. While OntoLex-Morph does not provide the level of detail of Mmoon, it defines elementary and reusable data structures for representing morphology as LLOD, and Mmoon is expected to serve as an inventory of morphological features in

this context. A desideratum in this regard is the wider application of the emerging OntoLex-Morph specifications to broad-scale morphological resources such as the UniMorph[8] and UDer,[9] and these are declared goals of the ongoing development of OntoLex-Morph specifications.

4.3. Pragmatics

Pragmatics studies the contribution of context to meaning and utilization of language in social interactions as well as the relationship between interacting interlocutors. To represent pragmatic information as LLOD, Pareja-Lora [73] extends the OntoLingAnnot annotation framework for morphological, syntactic, semantic, and discourse phenomena by an ontological conceptualization of pragmatics. To this end, pragmatic units are introduced to annotate text and dialogues in a way that they can interact with the other linguistic description levels, since every linguistic unit can have a pragmatic projection. For instance, `Apology`, `Begging`, and `Query` are instances of a `Speech Act` that in turn is a `Macroproposition`, a linguistic unit that follows from the aggregation of interrelated propositions from the `Discourse Level`. A `Macroproposition` is among others a subclass of a `Pragmateme`, the result of a text pragmatic analysis, and relations between pragmatemes are made explicit by way of a `Pragmatic Functional Unit`, such as a coherence relation. While the focus of this approach is on interoperability of linguistic description levels, the cited work exemplifies annotations in English without any reference to multilingual data.

In terms of discourse annotation, Chiarcos [74] proposes an extension of Ontologies of Linguistic Annotation (OLiA) [75] with a conceptualization of discourse features as found in major annotated corpora, e.g. Penn Discourse Treebank. To this end, the model introduces the classes `DiscourseCategory`, `DiscourseRelation` between instances of the former, and `DiscourseFeature` for annotations assigned to the former two. Thereby, the model allows for the representation of coreference and bridging, discourse structure and discourse relations, information structure (esp. topic and focus) and information status ((non-)given and (non-)salient). A predominant theory that guides the annotation scheme for discourse structure is the Rhetorical Structure Theory (RST), while discourse relations rely on Penn Discourse Treebank (PDTB). The OLiA discourse extensions build on earlier ontologies for discourse phenomena such as SemDok [76], an ontology of discourse relations used in a natural language generation system, and the Discourse Community of Practice Extensions [77] of the GOLD ontology [26], as well as on other efforts to standards discourse annotation schemas that originally used XML or domain-specific formats to model their taxonomies [78, 79]. The work on discourse annotation schemas stimulated the initiative on researching speaker attitude detection relaying on attitudinal discourse marker identification in the multilingual data. The speaker attitude detection is based on identifying discourse markers and the semantics of the discourse relations they introduce in text by using neural machine learning transformer models to ensure the interlinking of multilingual discourse markers [80].

Another line of research that broadly falls in the scope of pragmatics is the computational modelling of rhetorics, style and genre information by means of OWL ontologies [81–85]. At the moment, however, these are primarily conducted in the context of literary studies and less frequently applied to develop multilingual applications and thus beyond the scope of this article.

In terms of real-world applications, chatbots operating on knowledge graphs and other structured data have been described, as well as human language interfaces to ontologies or the use of ontology lexicalization techniques (e.g. [86, 87]. LINGVO [88], for instance, addresses the challenge of ranking knowledge graphs by their degree of multilinguality. While these technologies can benefit from and partially build on lexical data linked across multiple languages and thus have a multilingual dimension, the dimension and processing of discourse information is underrepresented in this line of research. A notable exception is the development and practical application of an OWL/DL ontology of discourse relations in the context of an NLG system by Bärenfänger et al. [76]. This general line of research from work on ontology-based parsing for symbolic natural language generation and deep syntactic proposed parsing around the time [89, 90], and is continued with limited intensity to this day [84, 91–94]. Overall,

[8] https://unimorph.github.io/, partially discussed in a LLOD context by [71]

[9] https://ufal.mff.cuni.cz/universal-derivations, the LLOD modelling of a related dataset for Latin is currently addressed in the context of the Linking Latin project [72].

however, the area is generally suffering from a lack of publicly available data sources compliant with Linguistic Linked Open Data formalisms. Instead, discourse-related data continues to be published in resource-, domain- or community-specific formats.

In an effort to address this issue, Chiarcos and Ionov [95] propose the formalization of discourse markers, such as *and*, *but*, and *though*, following the PDTB in the assumption that they trigger a discourse relation that connects an utterance with an element in the context. While this model represents an extension to OntoLex-Lemon, linking to the OLiA discourse extension is ensured. This last approach is particularly interesting within the context of this work as it not only addresses the capability to explore translation inferences, but extols the capability of querying discourse marker inventories across multiple natural languages. Valūnaitė Oleškevičienė et al. [80] in a preliminary approach propose to not only represent discourse markers as LLOD but to utilize them to detect speaker attitude with machine learning methods in text across natural languages.

While these approaches represent very valuable contributions to representing the pragmatic description level in LLOD resources, only the last approach explicitly addresses the potential that such modelling holds for multilingual and crosslingual pragmatics research. Thus, pragmatics represents one of the linguistic description levels with the lowest coverage in LLOD, in particular when it comes to multilingual LLOD. Apart from the work covered in this section, there is ample research in pragmatics from other perspectives not yet covered within the context of LLOD.

4.4. Lexicography

From a practical perspective, lexicography refers to the compilation, writing, and editing of dictionaries and other types of lexical resources. From a theoretical perspective, it relates to the study of lexeme features, such as syntagmatic and paradigmatic behaviour. A lexeme is coarsely defined as a set of inflected variants of a word.

Within the last years, a growing trend to publish lexical resources, including dictionaries, as linked data on the web could be observed. Bosque-Gil et al. [96] discusses the benefits of representing a lexicon in linked data, both from the macro-strucure (internal and external reusability of the elements in the lexicon, independence on the order of appearance of lexical entries and senses in cross-references, compatible onomasiological and semasiological views, etc); and the micro-structure (every lexicon element, i.e., lexical entry, sense, written form, etc. is a node in the graph, thus being a potential entry point in a LD dictionary). These and other advantages illustrate the difference between traditional electronic dictionaries, compiled with only the human as target, and creating them for both humans and computers, as it is the case of linked data dictionaries.

Some early works that used linked data to represent dictionary data comprise monolingual [97], bilingual [98], and multilingual [99] dictionaries, as well as diachronic [100], dialectal [101], and etymological ones [102]. Gracia [103] provides a description of two LOD resources consisting of bilingual dictionaries, i.e., Apertium RDF and Terminesp, the latter being described in Section 4.7. The data from these resources were converted into RDF by using the lemon model. Apertium[10] is an open-source machine translation platform containing over fifty bilingual dictionaries, also known as bidix. Out of them, 22 bilingual dictionaries were converted in a first effort and published in the LLOD cloud [98]. More recently, a new larger version of Apertium RDF was developed, by converting 53 bilingual Apertium dictionaries among 44 different languages into RDF. This new version was based on the more recent Ontolex-lemon model and it was used for cross-lingual model transfer in the Pharmaceutical domain [104]. Apertium RDF permitted the creation of a large unified RDF graph on the Web. The nodes of the graph are represented by the URIs of all the data elements from Apertium, e.g., linked lexical entries, translations. There are multiple ways to access and explore the graph, for example, by using SPARQL queries or dedicated search interfaces.

Based on the experience of the above referred works, Bosque-Gil et al. [105] identify a number of issues when converting information in a dictionary to OntoLex-Lemon, e.g. headwords may have different part-of-speeches. Also establishing translation relations between usage examples of words turned out challenging. The authors go on to propose a Lexicography Module to extend OntoLex-Lemon to resolve these issues. The specification of such a new module, called *lexicog*, was delivered as a W3C Community Group Report[11] and adopted by a number of initiatives such as K Dictionaries [106] and the Linking Latin project (LiLa) [107].

[10] https://www.apertium.org
[11] https://www.w3.org/2019/09/lexicog/

There has been a close collaboration between the recently finished projects Prêt-à-LLOD[12] and European Lexicographic Infrastructure (ELEXIS)[13] to provide use cases for linked data within the context of eLexicography [108]. Increasing interoperability of ELEXIS by means of linked data is, for instance, proposed in [109]. Relying on OntoLex-Lemon and other LLOD technologies, such as SKOS, the project shows how to port dictionaries to linked data (e.g. [110]).

The description level of lexicographic data is rather closed and quite well-covered with the proposed means approaches. However, several additional aspects, beyond purely lexicographic information that are covered in the following sections, still require further attention. For instance, handling etymological and diachronic information is still an evolving research topic.

4.5. Etymology and Diachronicity

Etymological information that provides details on word origins and histories is frequently a part of dictionaries. Thus, transforming dictionaries and lexical resources including etymological and diachronic information to LLD requires a means of adequately representing such information. Since OntoLex-Lemon is the predominant model for representing lexical information, Khan [111] proposed an Ontolex-lemon Etymological Extension (lemonETY) by linking etymological elements to `ontolex:LexicalEntry`. Before this extension proposal, both Gerard de Melo [112] and Ester Pantaleo [113] extracted the etymology information from the English Wiktionary edition and provided it as RDF using an ad-hoc modelling. The later is still available in the DBnary [114] dataset and a graphical application was built on top of this data for easy navigation in the etymology graph. Chiarcos and Sukhareva [115] convert dictionaries of historic language stages of Germanic languages and found the representation of original language abbreviations, especially hypothetical forms, e.g. Proto-Germanic, to be complicated, since LD and in particular OntoLex requires the assignment of ISO language codes. Such codes are not available for all historic languages and varieties.

In addition to word histories, it is important to enable a representation of historic languages and near-extinct languages with digital language equality and preservation of cultures in mind. Bellandi et al. [116] discuss how to represent a multilingual and multi-alphabetical Old Occitan medico-botanical lexicon in lemon and discuss an extension to multilingual settings, e.g. by extending `LexicalVariant` to `hasBilingualVariant`. Gillis-Webber and Tittel [36] investigate the representation of two near-extinct click languages of Southern Africa and the historic variety Old French as LD. The authors conclude that new language codes need to be created for language varieties and historic languages.

To truly assist in an inclusive approach to digital preservation of culture and cultural heritage, linguistic linked data should be able to accommodate all types of linguistic representation, i.e., written, spoken, and signed. Sign languages have received very little attention in LLOD, with very few exceptions, e.g. [117]. In this case, the topic goes beyond etymology and diachronicity, since the representation of sign languages as such already represents a blind spot. From a more etymological perspective, representing ancient signs, such as cuneiform signs, as LLOD should be considered. Homburg [118] proposes an extension of OntoLex-Lemon with paleocodes to this end, which requires an SVG representation among others.

Multimodal representations, as in the case of cuneiform signs and sign languages, represent one desideratum for the representation of linguistic description levels in multilingual linked data. Another major challenge in representing etymological and diachronic information as LLOD is the necessity to provide ISO language codes, which as a major desideratum should be extended to language varieties and historic languages in order to support digital language equality. Tittel and Gillis-Webber [119] extend this desideratum of additional language codes from a diachronic perspective to the dimension of diatopic, i.e., language varieties pertaining to a specific region. Diatopic-diachronic as well as diatopic-synchronic representation of languages is one description level that could benefit from more attention in LLOD.

[12] https://pret-a-llod.eu/
[13] https://elex.is/

4.6. Phonetics and Phonology

Phonetics studies the production and perception of speech sounds or equivalent representations, e.g. signs in sign language. Phonology studies how speech sounds, or equivalent representations, form patterns in a specific language or across languages.

The Phonetics Information Base and Lexicon (PHOIBLE) [34, 120] represents a phonological typology that ports disparate segment inventory databases to linked data to make them linguistically and computationally interoperable. Additionally, knowledge about distinctive features is added. Thus, PHOIBLE provides a research platform for segment and distinctive features across languages. A simple RDF model was created to link segments and languages, features and segments, and provide metadata for segment inventories.

Phonetics and phonology represents one of the least covered linguistic description levels in the LLOD, an assumption that is confirmed by the low coverage in our result set but also in other works on different LLOD linguistic description levels, e.g. [7]. A model to encode phonetic information has theoretically been proposed within the context of the General Ontology for Linguistic Description (GOLD) [26], which, to the best of our knowledge, has not been utilised to model data. Thus, one desideratum in this regard is to increase the phonological and phonetic coverage of languages in the LLOD.

4.7. Translation and Terminology

Translation refers to the explicit representation of equivalent words, terms or longer sequences across languages that derive from a translation process. In contrast, terminology describes the generally multilingual representation of equivalent domain-specific single- or multi-word terms across languages. Terminologies can represent translated terms or terms derived from parallel or comparable corpora.

Vila-Suero et al. [121] follow a similar path of addressing multilingual LD as Labra et al. [23] and identify three levels of multilinguality in a resource: the resource itself might be multilingual, the vocabulary to describe the resource might be mono- or multilingual, and a target dataset for enriching and linking might be mono- or multilingual. A use case on geo.linkeddata.es from the Spanish National Institute of Geography with metadata in several local languages is presented. While equally considering different aspects where multilingualism plays a role as in [23], the analysis is split into the method proposed by [122] for publishing LD: specification, modeling, generation, linking, publication, and exploitation.

Gracia et al. [123] propose an extension of lemon that builds on early work from Montiel-Ponsoda et al. [124] and introduces relations specific to modeling translations as linked data, such as `TranslationSource` and `TranslationTarget` as well as a set of categories to specify the type of translation, i.e., literal, cultural, lexical. This translation module is reused in other approaches, such as Zhishi.lemon [125] to represent links of translations from Chinese to other languages and resources. Such a translation module was the seed of the later *variation and translation* (vartrans) module of OntoLex-lemon,[14] which in addition to represent translations is able to represent any other type of lexico-semantic relation, including terminological variants. A more specific case is the representation of multilingual idioms, which LIDIOMS [126] introduces by means of ontolex and vartrans.

The DBnary dataset [114] draws on Wiktionary and provides vartrans relations for the subset of translations where source and target languages have their own lexicon, but introduced its own `dbnary:Translation` class when no target lexical entry is available. In this case, the translation is simply given as a string value, along with eventual context and usage notes.

León-Araúz and Faber [127] analyse the dynamic nature of terms and concepts from a pragmatic perspective and which challenges this raises for multilingual and cross-lingual settings. In terms of modelling, they utilise translation equivalents and context elements of OntoLex-Lemon. The main contribution is a detailed discussion of term variants from orthographic to diatopic and multi-dimensional facets of concepts as well as a detailed classification of terminological gaps and translation relations required to handle these gaps. Such relations are canonical translations, generic-specific translations, extensional translations, communicative translations, etc.

[14] https://www.w3.org/2016/05/ontolex/#variation-translation-vartrans

Early approaches to porting terminological information to linked data include Federmann et al. [128], where the authors present a new approach on the automated acquisition of multilingual terms for labels of ontologies in the financial domain from web stock exchange websites. This approach uses direct localisation/translation by searching candidate terms in various semi-structured multilingual web sources and repositories. Rule-based machine translation methods are used to extract terminology and work with under-resourced data extracted from multilingual websites. The final goal of this approach is to integrate the extracted terminology into Monnet [129] and TrendMiner [130] by transforming HTML into an XML-encoded multilingual terminology database or into the OntoLex-Lemon format. Multilingual terminologies available as LLOD described in Lewis [131] are among others IATE, EuroVov, TAUS, etc. More recently, Gracia [103] describes Terminesp,[15] a multilingual terminological database with Spanish technical terms. The majority of these terms also have translations in other languages, e.g., English, French, German. Terminesp was also published as a unified RDF graph [132]. Different to Apertium RDF, its structure is more a star-like graph, with Spanish in the centre.

Terme-à-LLOD [133] is a method of porting TermBase eXchange (TBX) resources, specifically as a use case IATE[16], to LLOD. To this end, a conversion to OntoLex-Lemon is proposed. An approach to automatically extract TBX terminologies including conceptual relations is proposed by Wachowiak et al. [134], where a direct RFD export is left for future work. Speranza et al. [135] show how Ontolex-Lemon can be used to add multilingual labels to an existing monolingual domain-specific terminological resource via identification of the relevant Wikipedia concepts.

This linguistic description level probably represents one of the better covered ones in the LLOD. In the vartrans model, there is even a relation type to foresee terminological relations to model term variant relations and lexico-semantic relations to represent relations between terminological units. However, in terminology it is common to propose a relation typology, which is a potential extension of this module that could be foreseen. Furthermore, in terminology and translation varying degrees of equivalence can be observed, ranging from overlapping characteristics to no equivalence. Currently, the main distinction is between full equivalence (ontological equivalence), partial equivalence and translatable in most contexts (translation), and minor equivalence in specific contexts (translatable as). Here a more fine-grained representation of equivalence with specific applications across languages could be of interest. In this context, it would equally be interesting to annotate the role cultural connotations play in the (lack of) equivalence since translation can be understood as a transcultural process, mediating between cultures. Explicitly annotating such cultural aspects for translations could open up interesting avenues for future translation-oriented research.

4.8. Approaches Considering Various Description Levels

While focused on the interdisciplinary exchange of theoretical and empirical findings on language acquisition research, Pareja-Lora et al. [136] address the need to integrate such data not only across disciplines but also across languages. Thus, they identify the necessity to describe and integrate language resources across different linguistic description levels, e.g. phonological information, morphological markings, syntactic differences, to perform cross-linguistic research. Cross-linguistic studies on language acquisition seek to identify commonalities and differences in developmental patterns across languages. The complexity of the data utilised for studying goes beyond linguistic description levels and extends to methodological and research design information, information about provenance (meta-data), multimedia representations of data (e.g. speech coding). All of these different dimensions should be captured and assimilated in order to allow a cross-resource analyses of research findings and data.

Two initiatives that have focused on representing language resources from different linguistic description levels, even though not directly related to LLOD but rather in the offline category of the language resource classification proposed by Lezcano et al. [137], are GrAF [138] and TEI [139]. Their LLOD counterparts are OntoLex-Lemon, Onto Media [140], MTE OLIA [141], ISOcat[17], among some other formats. Lezcano et al. [137] discuss several barriers to LR interoperability, which first of all relate to the phenomenon of a proliferation of representation formats

[15] https://aeter.org/terminesp/

[16] https://iate.europa.eu/

[17] ISOcat as such has been discontinued as an online inventory and has been succeeded by DatCatInfo, a repository of data categories, available at https:datcatinfo.net.

and standards and second to the underlying theories that require approaches seeking interoperability to consider several levels.

5. Resources and Their Use

Over time, LLOD resources have become available in all shapes and sizes and have been classified into different schemes, e.g. static vs. dynamic resources [142]. Static refers to inventories of data, whereas dynamic relates to creating new data, such as annotations. Furthermore, language resources can be monolingual or multilingual and relate to different domains or be domain-agnostic. To provide a structured overview of resources and their different uses, we rely on the typology of language resources in the LLOD cloud[18] as of May 2020, which are represented in the following and defined by Cimiano et al. [42]:

- **Corpora**: collection of language data, where either annotations and primary data are modelled in RDF or only annotations are provided as linked data
- **Lexicons and Dictionaries**: resources that focus on the general meaning of words and the structure of semantic concepts
- **Terminologies, Thesauri and Knowledge Bases**: resources that focus on vocabulary rather than linguistics and formalize semantic knowledge
- **Linguistic Resource Metadata**: metadata about language resources, including bibliographical data
- **Linguistic Data Categories**: metadata about linguistic terminology, including grammatical categories or language identifiers
- **Typological Databases**: collections of features and inventories of individual languages
- **Other**: resources that are not (yet) considered in the above classification

When it comes to using these resources, in this article we distinguish between linguistic data usage and LLOD use. Linguistic data usage refers to the scenario where data contained in an LLOD resource are re-used for some specific purpose, without benefiting from the fact that these data have been modelled as linked data, e.g. collecting strings from an LLOD lexicon. LLOD use refers to cases that truly benefit from the LLOD representation of language data and the full potential of Semantic Web technologies. Our focus in this article is on the LLOD use rather than linguistic data usage. LLOD is used in multiple ways which embrace linking multilingual wordnets, in digital humanities for multiple data models and vocabularies, and for discovery of language resources, and creating reusable multilingual annotated corpora.

Corpora. In recent years, and as an immediate result of the publication and reception of OntoLex-Lemon as the dominating community standard for this purpose, LLOD has been widely applied for lexical resources and is commonly seen as a building block to develop multilingual web technologies as already sketched by Buitelaar and Cimiano [143]. In the area of linguistic annotation, the situation is somewhat different, as several competing standards for annotation as LLOD have emerged that are both incompatible with each other, most prominently, Web Annotation [144] and the NLP Interchange Format NIF [145]. RDF versions of syntactically and semantically annotated corpora have been proposed as early as 2008, e.g. Burchardt et al. [146] porting the SALSA/TIGER corpus to an OWL-DL representation to provide a graph structure for flexible querying and consistency control. Other examples include the porting of the Austrian Baroque Corpus to LLOD [147] or porting a linguistic library to LLOD, including corpus information in OLiA [148]. Nevertheless, these standards are lacking the necessary data structures for morphology beyond the support for morphosyntax and inflectional morphology provided by terminology repositories such as ISOcat and OLiA.

In response to this, and specifically addressing the modelling of morphologically annotated corpora, Chiarcos and Ionov [65] introduced Ligt, an RDF vocabulary in accordance with classical interlinear glossed text (IGT). Based on established tools and formats such as FLEx and Toolbox [149], this is a minimal data model that allows encoding morphological segmentation, annotation and hierarchical structuring on all levels of morphology. Because Ligt is a

[18] https://lod-cloud.net/#subclouds

relatively novel contribution, it is not widely used yet, and it is primarily to be seen as a first step towards developing common specifications that address aspects of morphology in lexical resources and corpora (i.e., a synchronisation with OntoLex-Morph) on the one hand, and linguistic annotation in general (i.e., an extension or revision of Web Annotation or NIF to support morphological annotation) on the other hand.

One more recent example of converting annotations and primary data to the LLOD cloud is the conversion of the Tartar National Corpus "Tugan Tel" [150], making it possible to interlink the corpus with available Tatar linguistic resources, e.g. TatWordNet. In fact, a LLOD version of corpus data in general has the added benefit of providing interoperability with linguistic resources, be it corpora or other types [151]. One example from our result set is the semantic annotation project Open Access Database 'Adjective-Adverb Interfaces' in Romance, which links different heterogeneous multilingual corpora annotated morpho-syntactically and semantically in TEI/XML enriched with RDF [152].

POWLA [153] is a general formalism for interoperable representation of linguistic annotations through OWL/DL. In contrast to previous techniques in this area, POWLA is not restricted to a particular set of annotation layers; rather, it is meant to accommodate any kind of text-oriented annotation. Benefits of this type of representation are widely discussed, even for under-resourced languages (e.g. [37] for South African parallel corpora in our result set). Practical resources and applications in our result set are scarce and corpora are yet under-represented in the LLOD cloud in general. In particular, multilingual corpus annotations and interlinking multilingual corpus data is yet an underexplored area of research and practice.

Lexicons and Dictionaries. Language resources that provide elementary aspects of morphological information are manifold, as these aspects are already part of the OntoLex specification, but these primarily focus on morphosyntax and inflection. Racioppa and Declerck [63] show that LLOD technology allows to seamlessly merge traditional lexical resources, such as multilingual WordNet(s), with independently developed computational morphologies for the various languages, so that lexical entries can provide both sense information (from WordNet) and inflectional information (from language-specific morphologies). But, as specifications for the encoding of deeper morphological information in lexical resources are only emerging, only a limited set of lexical resources with rich morphological features are currently in existence, and these serve mainly as demonstrators of the respective vocabularies. As such, Klimek et al. [62] demonstrated the applicability of Multilingual Morpheme Ontology (MMoON) to encode morphology information for Hebrew.

Concerning wordnets of language resources, they include WordNet, Indo WordNet and Euro WordNet, which itself contains 76 wordnets in 47 languages[19]. The existing wordnets comprise over 200 languages, however, many of the wordnets are not complete or are not open. There were projects that aimed to link wordnets to external resources such as DBpedia/Wikipedia/Wiktionary. EuroWordNet is a multilingual database with wordnets for several European languages, which has been converted into RDF/OWL [154]. To achieve this conversion, the WordNet RDF-Schema was adapted to support the multilingual requirements of EuroWordNet by including OWL property conversion and domain extension. Furthermore, the RDF/OWL EuroWordNet resource was interlinked with both the pizza.owl and travel.owl by using a two-step approach that included the conversion of the domain ontologies OWL format to the EuroWordNet OWL format conversion and the integration of the converted data in the EuroWordNet hierarchy. Also, new relations were defined in RDF/OWL EuroWorNet in order to interlink and integrate the Hamburg Metaphor Database (HMD) and the Basic Multilingual Lexicon MEMODATA (BMD). The projects of BabelNet and UBY[20] attempted linking data in an automatic manner, whereas a semi-automatic mapping was proposed by McCrae et al. [155]. In order to manage the available WordNets, a new service called Collaborative Interlingual Index (CILI) has been created. It builds on standard LD vocabularies and the resource description framework (RDF) data model [15]. It should be observed that RDF is not fully embraced and the use of LMF and XML formats is still present in some cases.

Gillis-Webber [156] contributes to the important area of under-resourced languages by converting the English-Xhosa Dictionary for Nurses to RDF. This is particularly interesting, since it considers the representation of Click

[19] Even more wordnets are handled by the Global WordNet Association (globalwordnet.org).
[20] https://dkpro.github.io/dkpro-uby/

languages, requiring characters not typically included in a Roman alphabet. Taking a dynamic perspective on language data, particular emphasis is put on management of provenance and its related linked data generation.

Terminologies, Thesauri and Knowledge. Approaches that rely on SKOS as a data model for representing terminologies and thesauri range from AGROVOC to metadata. AGROVOC [157], a combination of agriculture and vocabulary, is a multilingual thesaurus of the Food and Agriculture Organisation (FAO) of the United Nations based on SKOS, currently available in up to 41 languages. The Linked Thesaurus Framework for the Environment, called LuSTRE [158], which also includes AGROVOC, is equally represented in SKOS. The Europeana project [32] relies on SKOS for its conceptual scheme and lexical semantic representation and then links literals found in metadata of paintings, books, newspapers, audio recordings, etc. to multilingual LLOD resources, such as GeoNames[21] and DBpedia[22].

An in-depth overview of the DBpedia knowledge base project is presented in [159]. DBpedia is a major interlinking LOD hub that extracts knowledge from more than 111 different language editions of Wikipedia. This knowledge base serves many purposes, and there are various applications and tools built around or applied to it. DBpedia project consists of several important components, i.e., the knowledge extraction framework, DBpedia ontology, and DBpedia Live. The knowledge extraction framework applies various extractors for translating sections of Wikipedia pages to RDF statements. The extraction is based on the community-curated DBpedia ontology, consisting of more than 320 classes. DBpedia Live provides live synchronization with Wikipedia with only small delays of at most a few minutes. In [160] the authors present a declarative approach implemented in a comprehensive open-source framework based on DBpedia to extract lexical-semantic resources from Wiktionary[23]. The main focus is on flexibility to the loose schema and configurability towards differing language-editions of Wiktionary. A declarative mediator/wrapper approach is achieved by using XML to extract the data from different pages. The extracted data is as fine granular as the source data in Wiktionary and additionally follows the lemon model. Closely related is the idea to create a Multilingual Wikipedia Bitaxonomy (MultiWiBi) introduced in [161].

In [162], the authors present an overview of large-scale multilingual parallel language resources made publicly available by the European Commission (EC) and different European Union (EU) organisations with the aim to clarify what the similarities and differences between the various resources are and what they can be used for. The work focuses on 7 full-text corpora resources that cover all 24 official EU languages as well as a variety of non-EU languages: JRC-Acquis [163], DGT-Acquis and Digital Corpus of the European Parliament (DCEP) [164], the translation memories DGT-TM [165], ECDC-TM and EAC-TM, and the document collection accompanying the multi-label categorisation software JRC EuroVoc Indexer (JEX) [166]. These resources are made publicly and freely available online through the Europe Media Monitor (EMM) [167] family of applications developed by the Joint Research Centre (JRC) - EC's in-house science service.

One resource in the category of knowledge bases is the Semantic Quran [64], a multilingual RDF representations of translations of the Quran. Building on an ontology specifically designed for this resource, the dataset encompasses 43 languages including some of the most under-represented in the LLOD cloud, such as Arabic, Amharic and Amazigh. The format is compatible with the NIF format and eases application scenarios, such as data retrieval for training NLP tools or linguistic research including morpho-syntactic aspects due to explicit representation of morpho-syntactic information.

Another endeavour to link a knowledge base with the Linked Data cloud is described in the project of integrating EcoLexicon, which is a multilingual (Spanish, English, German, Modern Greek, Russian, French and Dutch) terminological knowledge base, into DBpedia and GeoNames. The project is based on 'linking legacy systems (RDB stored information) with an ontological system' [168]. Also Web technologies are applied in Digital Humanities including their application in APIs, NoSQL databases, and database integration as well as terminology management. Linked Open Data is increasingly applied in digital humanities for LOD resources (prosopographical databases, gazetteers, citation services) and in other projects and applications. The vocabularies created by the linked data movement are broadly adopted in digital humanities and used for terminology integration over the distributed data

[21] https://www.geonames.org/
[22] https://www.dbpedia.org/
[23] https://en.wiktionary.org/wiki/semantic

collections, for example, SKOS, CIDOC-DRM and CTS. The metadata vocabulary in the GLAM provides data on galleries, libraries, archives and museums; there is also Linked Geo Data. A project of collecting, digitising and tagging Geolinguistic data of Cimbrian dialect varieties also adopted the LOD approach to make the dataset interoperable and available to other researchers and projects [169].

From the administrative and legal domain, a major LLOD resource is the multilingual EuroVoc vocabulary from the European Commission published in SKOS [170]. A more comprehensive initiative to port to and interlink legal language resources in the LLOD cloud was proposed by Martín-Chozas et al. [171]. Their approach includes the porting of existing resources, such as German Labour Law Thesaurus and JuriVoc, to RDF as well as the creation of new resources drawing from automated term extraction and existing legal language corpora. Moreover, LOD has become relevant for accessibility and transparency of government data publication worldwide. Researchers of the World Wide Web Consortium [172] have designed best management practices for publication and interlinking high-quality government data via the RDF and SPARQL. It also should be stressed that the popular TEI data model used in digital humanities can be made compatible with RDF. From a different angle, Gromann [173] presents a vision of joining Neural Language Models (NLM) and LLOD towards a multilingual, transcultural, and multimodal information access. Different linguistic description levels are not considered explicitly, however, methods and application scenarios for all three dimensions are provided. In terms of the multilingual aspect, such a work proposes uniting different application scenarios of Neural Machine Translation (NMT) and LLOD, e.g. translating LLD contents, learning structured knowledge with NMT, or building reasoning on NMT, and NLM-based ontology alignment.

From a different perspective, in [174] a method is proposed that employs the use of Machine Translation techniques (e.g., Bing Translator[24]) to identify links between documents (i.e., thesauri) written in different languages. Another interesting approach is the QLAD challenge, which has the objective to evaluate natural-language based question answering interfaces to linked data sources, i.e., sources that are characterized by their large scale, openness, heterogeneity, and varying levels of quality [175].

Linguistic Resource Metadata. Available resources per type and/or language can be discovered using repositories of language resources with detailed linguistic resource metadata which are maintained by dedicated organisations, such as META-SHARE[25] or the CLARIN[26] project's Virtual Language Observatory (VLO)[27]. Such moderated repositories enables to ensure high-quality metadata entered and edited by experts, however, limiting the coverage. The other method is a collaborative approach, for example, the LRE Map[28] or Datahub.io[29], which allow anyone to publish language resource metadata increasing the coverage but decreasing the control over the quality. An approach to reconcile linguistic resource metadata from all these repositories as linked data in a single interface has been presented in the form of LingHub[30] [176, 177].

Linguistic Data Categories. Chiarcos and Sukhareva [28] present the development of the Ontologies of Linguistic Annotation (OLiA) [75] since 2006, which provide comprehensive annotation terminology for linguistic phenomena. OLiA, with a modular architecture of OWL2/DL ontologies, includes four different ontologies: the OLiA reference model, which describes the common terminology used by different annotation schemes; OLiA annotation models, which formalise annotation schemes and tagsets; a linking model, which establishes relationships between the concepts/ properties in the annotation model and reference model; and external reference models, which are terminologies repositories that are integrated in OWL2/DL. OLiA compiles annotation terminology, and works as an interlingua between the annotation schemes of different linguistic resources and the external reference models to which it is linked. OLiA provides links to other existing linguistic data category repositories, such as the General Ontology of Linguistic Description (GOLD), ISOcat, OntoTag and Typological Database System (TDS). Chiarcos

[24] ttps://translator.microsoft.com/
[25] http://www.meta-share.org/
[26] https://www.clarin.eu/
[27] https://vlo.clarin.eu/
[28] https://lremap.elra.info/
[29] https://datahub.io/. Unfortunatelly Datahub changed its business model and discontinued their free online repository. The datasets that were previously hosted there were transferred to https://old.datahub.io/
[30] https://linghub.org/

and Sukhareva [28] also document different application scenarios of OLiA, such as interoperable corpus queries, interoperable information processing in NLP pipelines, and ontology-based NLP.

One more project converted the semantic resource Thompson Motif index (TMI) of folk-literature into LLOD based on porting lexical resources provided in Wiktionary to a standardised representation, with the aim to support 'semi-automatic translation of TMI' and 'the automatic detection and semantic annotation of motifs in literary work, across genres and languages' [178]. The multilingual value of this project is reflected in an attempt to enrich TMI, which contains labels in English only, by labels in other languages, namely, German and Hungarian.

Typological Databases. One very early approach to address typological queries across languages building on linked data principles is the "Typology Tool" (TYTO) [179], which seems to not be available anymore. A strategy targeted at less-resourced languages integrates the catalog for linguistic data categories Glottolog/Langdoc with lexical-semantic resources of the Autom ted Similarity Judgment Program (ASJP) [180]. This approach seeks to represent genetic relatedness between languages based on their lexical distance. In a later work, Nordhoff [67] harvests and interlinks glosses and metadata from an archive of endangered language to provide this information in 280 low-resource languages as LLOD building on Ligt [65]. A similar approach has recently been taken by Ionov [181] in converting the Atlas of Pidgin and Creole Language Structures (APiCS) IGT dataset to Ligt.

An additional model in our result set of publications is the Model for Language Annotation (MoLA) [182]. MoLA provides an RDF vocabulary for language annotation that permits the definition of custom language tags and their association with a time period and region. Furthermore, our result set contained the Cross-Linguistic Data Formats (CLDF) building on the CLLD project [183] that represents data types for language typologies. An example of a typological database modelled with CLDF is the representation of languages or rather languoids inspired from GLOTTOLOG, which models parameters that can be compared across languages, values of these parameters, and source referring to the primary source of data collection [184]. It further specifies the CLDF modules, e.g. wordlists, parallel texts, etc., and CLDF components, e.g. cognates, functional equivalents, etc. This format has been applied to various resources, including a database of cross-linguistic co-lexifications in more than 3,000 language varieties with the objective to analyse cross-linguistic polysemies [185] and the phylogenetic methods to analyse the ancestry of Sino-Tibetan [186].

Other. According to the LLOD cloud typification, a very large, multilingual resource that has been classified as "Other" is BabelNet [187], initially based on data from both WordNets and Wikipedia. BabelNet links information from complementary resources. On the one hand, highly structured lexical databases, for example, WordNet and the like [188], containing lots of lexical semantic relations of different kinds between words (word senses) and, on the other hand, encyclopedic information from Wikipedia (Named Entities) are jointly accessible in BabelNet [187]. Interlinking both types of resources mentioned above makes BabelNet a useful LL(O)D resource fostering integration, reuse and interoperability of other resources, both resources that could be included in versions of BabelNet and resources/tools that can be built making use of BabelNet. The integration and interoperability could be illustrated by the use of such tools like Semantic Textual Similarity: how similar two texts are at the semantic level, *in se* independent of the language used in these texts or (Neural) Machine Translation, making use of concepts in BabelNet, especially for low resourced languages. In a later stage other resources were added, like OmegaWiki and GeoNames. BabelNet is provided as a stand-alone resource with its own Java API, a SPARQL endpoint and a linked data interface as part of the LLOD cloud.[31]

Another resource that is not yet classified is the publication of Joint Research Center (JRC)-Names resource as linked data using OntoLex to address the problem of identifying name variants of entities found in news media worldwide, within and across many languages [189]. The JRC-Names data originate from real-life multilingual texts, containing useful, complementary name variants.

[31] The last available version of BabelNet as LLOD is 3.6, released on February 2016. Later updates of BabelNet (the last one is v5, at the time of writing this), do not contain updates of the linked data version.

6. Challenges

Despite its rising popularity and recognition of its usefulness by different disciplines, the LLOD Infrastructure has some new [108, 190] and old [12] challenges to overcome. As a result of our systematic study, and also based on our own experience, we analyse in this section a number of such challenges to be addressed in order to bring LLOD to its full potential for representing and linking multilingual language data across linguistic levels. Notice, though, that some of such challenges are common to LD in general (e.g. sustainability), however, we do not want to miss the opportunity to refer to them here because they are also crucial for the LLOD community. Other issues related to language resources or linguistic data in general but not so much specific to LD or LLOD (e.g. legal issues, ownership, data protection [21]) are out of the scope of this section.

6.1. Entry Barriers to the Technology

One of the main challenges concerns the use of the LLOD Infrastructure by researchers unacquainted with its framework. In fact, as any other technique in an early adoption stage, the use and application of LD requires to face a steep learning curve (RDF, OWL, SPARQL, specific models such as OntoLex-Lemon, etc.). Furthermore, new adopters will need a certain technical support to setup the appropriate infrastructure, which may vary depending on their needs (from simple storage of RDF dumps to fully-fledged triple stores with de-referenceable mechanisms).

Another challenge results from the amount of language resources that are available, which increases the complexity of issues related to interoperability. In fact, once a resource in the LLOD cloud is discovered, its access and exploitation is not always straightforward. Additionally, the presence of abandoned resources and broken links in the LLOD cloud might be a discouraging experience for newcomers.

For these reasons, not only the development of tools, standards, and overall research, but also an investment on education by means of training schools and courses, are crucial for the maintenance and advancement of the LLOD infrastructure, by the growing LLOD community. In that respect, ongoing research projects and networks, and the activities of several WC3 community groups, are progressing in that direction. For instance, NexusLinguarum[32] is organising a series of training schools around the topic of linguistic linked data, and has supported a number of tutorials and seminars on this topic. Additionally, Linghub, developed in the context of the LIDER[33] and Prêt-à-LLOD[34] projects, aims at alleviating the issue of discoverability and reusability of language resources [176], by indexing a large amount of language resources metadata in a way that can be easily exploited by software agents as well as by humans.

There is, however, a larger need of visual interfaces and working environments to deal with LLOD (frameworks such as VocBench [191] are a step in the right direction), as well as tools and infrastructures for an easier deployment of (linguistic) semantic data on the Web. Previous efforts like the *lemon source* framework [192] that targeted the collaboration of experts and non-experts in a collaborative semantic editing environment for linked lexical data, similar to a wiki, were highly appreciated, however, unfortunately discontinued. This again shows the high need for persistence of LLOD tools and technologies. Additionally, the design of multilingual user interfaces poses a challenge [31].

Researchers and practitioners that specialise on specific linguistic description levels and actively generate linguistic resources covering one or more linguistic description level are not necessarily LLOD-savvy. Lowering the LLOD entry barrier is in the interest of the LLOD community as well as of these researchers and practitioners. For the former, it is important to increase the coverage especially of yet under-represented linguistic description levels, such as phonetics and phonology, pragmatics, dialogue, sign languages, and diatopic representations. For the latter, it is of interest to maximise the re-use and interoperability of their often manually curated resources.

[32] https://nexuslinguarum.eu/
[33] http://lider-project.eu/
[34] https://pret-a-llod.eu/

6.2. Sustainability

As it has been recently reported in several fora[35] and scientific papers [193], there is a need of sustainable hosting solutions for the RDF data exposed as linked data on the Web. The main issues, which are common not only to LLOD but to LOD in general, are:

1. Data consumers may want content negotiation mechanisms and server side infrastructure (triple store + SPARQL endpoints). This can be a burden on the host/provider.
2. Alternatively, the burden can be put on data consumers, if they need to download and locally process RDF data dumps.

The challenge here is how to balance efforts between data provider, data consumer and data host. Focusing on the federation and queryability of linked data resources, a scenario that is ideal from the perspective of the user would be if the host can expose the data via a SPARQL endpoint – which can be directly queried by a client without setting up local infrastructure. On the other hand, real-world infrastructures currently allow only to deposit data *as files* with the media types plain/text (plain text) or application/octet-stream (arbitrary binary data). In order to use this data as RDF, an application needs to guess the correct format, and in many cases, it requires to download all data first and set up a local query engine. One compromise between both extremes is to deposit data as uncompressed files *with appropriate RDF-compliant media types* (e.g., text/turtle, application/ld+json, etc.), with a small additional burden on data provider and host to indicate the proper media type, e.g., by means of content negotiation) [193]. Then, the data can just be imported into an RDF data base (or a SPARQL web service) by means of the SPARQL keywords `LOAD` or `FROM`. On a technical level, some other intermediate solutions have been proposed, like:

- Linked data Fragments[36] is an effort to redistribute the load between clients and servers by means of the Triple Pattern Fragments [194].
- SPARQLer[37] is a web service that allows to run queries against external data sets that can be consulted using the SPARQL `FROM` key word. SPARQLer is just a blank installation of Apache Jena[38] with permissions granted to eliminate the need for a user to set up a local RDF database.
- RDF-HDT is a community standard for binary compressed RDF data that can be directly queried by means of SPARQL [195]. HDT requires to download external data, but does not require to set up a local SPARQL end point.

More powerful support and infrastructures are, however, still needed. Something analogous to www.wordpress.org for web sites, but for small linked data providers. Some steps in this direction are Databus[39], TriplyDB[40], and Semantic media wiki[41]. We consider that larger infrastructures, like the European Language Grid[42] (ELG) or CLARIN[43] can play an active and important role here.

6.3. Coverage of Current Representation Models

In order to lower the entry barrier to the LLOD cloud and enable researchers and practitioners to publish their data as linked data with ease, a representation mechanism for the respective data is a paramount prerequisite. While this is the case for most of the linguistic description levels that we identified in this survey article, not all aspects of linguistic research are yet represented. One level that encompasses more facets in linguistic research than LLOD

[35] https://www.clarin.eu/event/2021/clarin-cafe-linguistic-linked-data
[36] https://linkeddatafragments.org/
[37] http://www.sparql.org/
[38] https://jena.apache.org/
[39] https://databus.dbpedia.org/
[40] https://triply.cc/
[41] https://www.semantic-mediawiki.org/
[42] https://www.european-language-grid.eu/
[43] https://www.clarin.eu/

representations currently provide is phonetics and phonology. PHOIBLE 2.0[44] provides a very large cross-linguistic inventory of phonemes in more than 2.000 languages, however, it is one of the few LLOD models for this description level available and many areas from socio-phonetics to phonetics in language acquisition might require a dedicated representation. Areas such as sign phonetics from a multilingual perspective, not solely focusing on a specific sign language, and representing sign languages as LLOD resources in general is yet to be explored systematically. Regarding the level of pragmatics, there are some models, such as the OLiA discourse extension, that focus on representing dialogue structure, however, also this linguistic research field has more to offer, e.g. speaker attitude, turn taking, etc.

Another important aspect of representing linguistic data as linked data is the ease to move across and between distinct description levels. Fortunately, interoperability is one of the key assets of the LLOD concept. One predominant approach of the LLOD community that becomes evident in this survey is the extension of existing representation models with dedicated modules for specific levels. For instance, numerous extensions to OntoLex-Lemon and OLiA provide a communal base representation to which to link specific information, e.g. phonetic features and morpho-syntactic annotations across languages. Models with different theoretical underpinnings can equally be jointly explored by means of their linked representation in the LLOD cloud. However, this brings us back to the ease of access to LLOD resources, which is a requirement to be attractive to a wide audience. Only then is it feasible to explore cross-disciplinary linguistic research in multiple natural languages.

When it comes to specific language resources, especially corpora, formalisms such as POWLA have been proposed a decade ago, but still very few primary corpus data or corpus metadata have been published in the LLOD cloud. This raises the question whether there is a need to extol the virtues of querying, consistency controlling, and linking such data, also to other types of resources and across languages, more explicitly or whether the entry barriers to the LLOD cloud and/or representation models is too high for providers of such data. Within the COST Action NexusLinguarum[45] there has been an initiative to collect feedback from corpus providers on the use of LLOD in this context. Despite the results not being conclusive yet, they indicate that large national corpus providers tend to be reluctant to utilise linked data, if they had even heard about it, stating that resources tend to be unstable (without automatic redirects if a resource fails), that it is hard to integrate linked data with current machine learning methods, and that there is a lack of tutorials for LLOD Infrastructures. These arguments rather suggest that the reluctance to publish corpora as linked data is more an issue of LOD Infrastructure, which needs to become more stable, easy-to-use, and ideally integrated with state-of-the-art machine learning methods, than with proposed representation models. Nevertheless, this survey article shows some representation models have been taken up more vibrantly than others, which might not necessarily allow conclusions about the model itself but rather constitutes a call to the LLOD community to more closely interact and collaborate with communities that curate multilingual data. For instance, strong showcases of performing multilingual linguistic research on an easily accessible LLOD Infrastructure might help the case.

6.4. Metadata

Another remarkable issue when publishing LRs on the Web is that their metadata is scattered across the different language repositories, which makes it problematic to ensure effective search procedures across the repositories. Furthermore, there are different standards adopted for different repositories, which makes the data accessibility and linking problematic. There are also difficulties in harmonising metadata from different repositories in order to provide a single point of access to search for relevant language resources across repositories.

Actually, linked data provides suitable mechanisms to solve such issues. In this regard, we advocate for an increased use of agreed vocabularies for LRs metadata description, such as the Meta-Share OWL ontology [196]. An example of the use of the Meta-Share ontology can be found in the aforementioned LingHub service. Other types of metadata that might be of interest for the LLOD cloud is the Information Coding Classification (ICC) [197], or the licensing information in machine-understandable ways [198].

[44] https://phoible.org/
[45] https://nexuslinguarum.eu/

Besides metadata for the description of language resources, metadata for the development of particular use cases in linguistics also poses interesting challenges. For instance, as reported by Blume et al. [199] the use of LOD for research on multilingualism, particularly on language acquisition, require a set of very different metadata to characterise multilingual speakers that currently are not present in the LLOD cloud, to account for psychological and sociological factors, competence being evaluated, language speaker's acquisition history, among many other features. In fact, means to represent information on discourse structures and discourse relations in a multilingual setting and pragmatics in general is currently poorly represented in LLOD, as are phonetics and phonology. One especially challenging aspect within the context of LLOD is that all these metadata need to be linked to the participant in a specific study rather than to a language resource or a data repository. Thereby, LLOD could support the development of meta-analysis studies, e.g. to analyse the development of a specific grammatical element across studies. Furthermore, as studies on translation inferences in general and in relation to pragmatics have shown, the potential to query data inventories in a structured manner with a specific research question in mind across languages, potentially even from a diachronic perspective, open up entirely new research avenues for different linguistic branches. For phonology, for instance, such interlinking holds the potential to analyse speech patterns across a large number of languages and representation modes.

6.5. Cross-Lingual Linking

Interlinking multilingual resources is not straightforward since when entities are described in different natural languages, string similarity measures cannot be applied directly. This task poses several challenges [200]: (1) the structure of graphs can be different and the structure-based techniques will not be of much help; and (2) even if the structures are similar to one another, the properties themselves and their values are expressed in different natural languages. In this regard, even though a Natural Language Processing (NLP) approach is adopted, the performance of the method may depend on the amount of text and discriminative power of labels [29, 30].

From the perspective of conceptualisation, another issues arise in the linking task [201]: a) conceptualisation mismatches due to language and cultural discrepancies; b) conceptualisation mismatches due to the perspectives from which the same domain is approached; or even c) different levels of granularity in the conceptualisation. Despite the recent advancements in the field, all the referred issues remain valid and give room for further research.

Another remarkable challenge is the need of benchmarks to support the evaluation of methods and algorithms on cross-lingual linking, in a Semantic Web context. Current efforts in that direction are the Multifarm [202] track, which is part of the periodic Ontology Alignment Evaluation Initiative (OAEI)[46], and the Translation Inference Across Dictionaries (TIAD)[47] shared task [203]. The Multifarm dataset is composed of the alignments among seven ontologies of the Conference domain, translated into eight different languages, thus resulting on 45 different language pairs that serve as gold standard for cross-lingual ontology matching systems. Despite its obvious interest, this dataset only covers one specific domain. More domains and languages would be necessary to further stimulate the progress in the field. Additionally, the TIAD task has being beneficial and led to progress in the field of cross-lingual linking. However, this is specific to a concrete task, which is bilingual lexicon induction, and measures performance among three language pairs (French, English, Portuguese) only. A broader language coverage and the extension of this idea to similar tasks involving cross-lingual link discovery would be also beneficial.

6.6. Under-Resourced Languages

There are a number of works in the scientific literature that clearly illustrates the potential and usefulness of LLOD for under-resourced languages [34–37, 204]. There are some remaining open issues, though, like the necessity of modelling under-resourced languages that are very morphologically rich, which contrasts with the still low adoption of LLOD at the morphological level. A second remarkable issue, as pointed out by Gillis-Webber and Tittle [36] is the current limitations of language tags when dealing with very specific language variants or dialects. The latter is, however, not an LLOD-specific issue, but something broader that involved internationalisation of the Web at a larger

[46] http://oaei.ontologymatching.org/

[47] See latest campaign description at https://tiad2022.unizar.es/

scale. Nevertheless, potential solutions to that issue might come in linked data native ways following the example of lines of works such as Lexvo.org [205], a database that brings information about languages, words, characters, and other human language-related entities in a linked data format.

6.7. Multilinguality

The Semantic Web in general, and linked data in particular, has been repeatedly identified as a core technology to overcome language barriers on the Web [12, 206], since it has mechanisms to represent, traverse, and integrate, data in different languages, mediated by a common ontological layer. However, has LLOD really helped in making the Semantic Web more multilingual? Studies indicate that the number of language tags used in the Semantic Web increased, but the dominance of English never stopped [207, 208]. New studies should maybe take a more up to date snapshot of the current status of the LOD cloud with respect to the availability of data in different languages.

In terms of comparison of the LLOD cloud and the broader LOD one, one wonders if LLOD is more "multilingual" than the general LOD. The current availability of linguistic data in the LLOD in terms of languages needs a more systematically exploration. There is also a need to focus on the coverage and details on the granularity of available data (lexical entries / links to other languages through translation of common referents / availability of data from the different linguistic description levels / etc.). An "observatory" would be needed to measure the quality and evolution of linguistic data along such dimensions.

7. Towards an Ideal Ecosystem for LLOD

In a previous analysis, one decade ago, Gracia et al. [12] studied the challenges posed by the so-called Multilingual Web of Data and proposed a roadmap towards its full realisation. In a first stage, they proposed the development of new (lightweight) representation models along with simple techniques for ontology localisation, cross-lingual querying and linking. The idea was to ensure early adoption of LLOD and provide the required incentives for the development of more complex infrastructures in future stages. In a second stage, semantic search engines might index multilingual lexical information available on the Web and support answering ad-hoc queries in any language. More complex models and services would be developed in this second stage, supporting cross-lingual natural language processing applications requiring deeper multilingual lexical knowledge. Finally, the third stage would be more user-centered, with people more motivated to provide multilingual lexical information. An ecosystem of services would be available for cross-language querying, on-demand translation, cross-lingual mappings, etc. Search engines might be able to process natural language questions in any language and adapt their result presentation to conventions of the linguistic and cultural community to which the user belongs.

As our literature analysis attests, there has been substantial progress in the field over the last ten years. However, this progress did not always move in the direction predicted in the mentioned roadmap. Some goals have been accomplished, to judge from the emergence of new models (e.g., lexicog [105]) and updated versions of other well-stablished ones (e.g., lemon [41]), as well as the (still moderate) progress in cross-lingual link inference (e.g., TIAD campaign [203]). However, the roadmap envisioned a more central role for the final Web user, more aware of the incentives and rewards that publishing linguistic information as LD should bring. We are still far from that. Recent progress has been achieved mainly in academic contexts, for specialised studies with specialised linguistic data. This is not bad in itself, of course, and there are very successful stories in the application of LLOD for linguistic research (e.g., the LiLa[48] project [72]). However, some pieces are still missing for a larger uptake of the LLOD technologies. For instance, a major role of semantic search engines, as envisioned in the 2012 roadmap, or a higher level of infrastructural/sustainability support, as reported in Section 6.

In the rest of this section, we propose a new roadmap with the next steps that the community might take to address the challenges reported in Section 6, in order to attain an ecosystem of truly interoperable linguistic data on the Web, multilingual in nature, across different linguistic levels. These steps are not intended to be sequential and can overlap.

[48] https://lila-erc.eu/

1. Step I. More robust and sustainable open infrastructures should be in place, to support small and medium scale data providers who cannot afford their own hosting infrastructure. Since the technology is already in place, this is a matter of promoting its adoption and carrying out new national and international LD projects with a clear focus on infrastructure development. In parallel, more educational efforts are needed to make the advantages of LLOD visible to a new generation of researchers and practitioners. While this step is a general LOD issue, it is of crucial importance to achieve a highly Multilingual LLOD cloud as this necessarily requires publishing many datasets of varying size and language coverage from many data publishers who cannot afford their on-premise infrastructure.
2. Step II. New models, along with new systems for RDF generation and linking, will be developed to cover linguistic description levels currently under-represented in the LLOD cloud. This will enable truly cross-disciplinary linguistic research in multiple natural languages, at Web scale.
3. Step III. Development of an "observatory" to measure the quality and evolution of linguistic data on the Web along several dimensions (language, linguistic level, usage, etc.). Stable metadata models and repositories will be in place, with the ultimate aim of not only discovering relevant language resources, but really accessing to their data and enabling their direct re-use and inter-operation. Metadata models are of tremendous importance in Semantic Web and LOD in general. Their usage are, however, mainly disregarded in the NLP community.[49] This step is the key towards usages where the required resources would be automatically discovered and used in the LLOD, rather than fixed (and usually imported) at development time.
4. Step IV. Massive population of the LLOD cloud with the maximum possible number of languages (thousands better than hundreds) and resources. That will create a critical mass of data to be eventually exploited by final language applications. This should cut the vicious circle resulting in lack of data caused by lack of exploitation opportunities and vice-versa.
5. Step V. Development of a fully fledged family of services for easy upload and integration of multilingual linguistic data on the Web, language independent access and querying of linguistic data, and seamless integration of such a data with NLP services and tools. That will include also user interfaces for browsing/editing linked data.

8. Conclusion

This systematic survey on the status of multilinguality and LLOD that is built on the PRISMA method aims to provide an overview of available representation models, resources, and approaches for and across different linguistic description levels, pointing out existing challenges and gaps. It contributes (i) a guide on the state-of-the-art for researchers and practitioners interested in exposing their linguistic data as LLOD with a focus on available approaches for specific linguistic description levels. Furthermore, it (ii) identifies open challenges and gaps in the support of specific linguistic description levels across multilingual LLOD resources. For the LLOD community, this survey presents a report on where to direct future joint efforts towards multilinguality and LLOD. Among the identified description levels, phonetics, phonology, pragmatics and discourse structures have turned out to be least explored, correspondingly wanting in representation means. From a resource perspective, available formalisms have not necessarily resulted in a wide publication of linguistic data, e.g. corpora and typological databases are quite under-represented in the LLOD cloud. Finally, (iii) we present a solid basis for future best practices on how to represent, model, and link different linguistic description levels in a truly multilingual LLOD cloud. To this end, this article proposes an ideal ecosystem, that is, a step by step roadmap to linguistically-rich multilingual LLOD, which addresses general LLOD challenges as much as LLOD challenges particular to multilinguality and LLOD.

Results of this article indicate that most individual description levels are well represented and that for most types of language resources examples exist, however, they also suggest that the key asset of the LLOD representation of interoperability should be more extensively explored for **cross-disciplinary linguistic research** across natural

[49] Indeed Ducel et al. [209] recently showed that around 32% of ACL research papers do not mention the language that is studied while they should have.

languages, which represents another future avenue of research. To this end, the presented survey identified a number of key challenges of multilinguality and LLOD.

One of the first and foremost challenges has been and still is **lowering the entry barrier to LLOD** and LOD. Hence, it is highly important to increase ease of access by providing graphical user interfaces with a high degree of usability and representation in as well as support of multiple languages, considering different linguistic description levels. First solutions, such as VocBench, have been proposed in this direction, however, a closer collaboration with linguists and computational linguists is required to provide solutions that are truly usable across disciplines. Some first efforts to increase this cross-disciplinary collaboration on LLOD can be observed, such as the COST action NexusLinguarum, which also provides training schools, another important ingredient for lowering the entry boundary. Nevertheless, any of these efforts depends on solving the central challenges of **sustainability**, that is, consistent availability of support and a stable infrastructure for LLOD. As a mostly research-derived initiative, ways of ensuring a persistent publication method of language resources and their use cases are crucial.

In terms of **representing different linguistic description levels**, many representation models have been proposed, however, not necessarily for all levels or to the degree needed to cover all aspects, e.g. of morphologically-rich under-resourced languages. Thus, besides the need for a kind of "observatory" to monitor the development of the LLOD cloud, tracking and actively promoting the uptake of models might accelerate the proliferation of linguistic description levels and language resources as LLOD. For only models that are actually used can be regarded as truly validated as a means of representation, whereby the call for more collaboration with language resource providers comes into play again. This is equally true for **metadata** initiatives, where some interoperable solutions for language resources have been provided, but not for all linguistic description levels and especially not for all potential features or characteristics for specific use cases. For instance, use cases related to discourse structures might need to represent demographic, social or psychological characteristics of speakers. Finally, even though this paper focuses on multilinguality, challenges pertaining to **cross-lingual linking** should be considered, which mainly concern different theoretical underpinnings, graph structures, and levels of granularity of LLOD language resources. A strong benchmark for cross-lingual linking might contribute to the development of this area.

Lastly, we have envisaged an ideal ecosystem for LLOD in the form of an open, multilingual and semantically interconnected linguistic data environment that facilitates access and interoperability, offering features that are universal, transdisciplinary, transnational, and translingual.

Acknowledgments

This article is based upon work from COST Action NexusLinguarum – European network for Web-centered linguistic data science (CA18209), supported by COST (European Cooperation in Science and Technology). It has been also partially supported by the Spanish project PID2020-113903RB-I00 (AEI/FEDER, UE), by DGA/FEDER, and by the *Agencia Estatal de Investigación* of the Spanish Ministry of Economy and Competitiveness and the European Social Fund through the "Ramón y Cajal" program (RYC2019-028112-I).

References

[1] J.J. Gumperz and S.C. Levenson, Rethinking Linguistic Relativity, *Current Anthropology* **32**(5) (1991), 613–623. doi:doi/10.1086/204009.

[2] P. Cimiano, C. Chiarcos, J.P. McCrae and J. Gracia, Link Representation and Discovery, in: *Linguistic Linked Data*, Springer, 2020, pp. 181–196. doi:0.1007/978-3-030-30225-2_10.

[3] G. Budin and A.K. Melby, Accessibility of Multilingual Terminological Resources - Current Problems and Prospects for the Future, in: *Proceedings of the Second International Conference on Language Resources and Evaluation (LREC'00)*, M. Gavrilidou, G. Carayannis, S. Markantonatou, S. Piperidis and G. Stainhauer, eds, European Language Resources Association (ELRA), Athens, Greece, 2000.

[4] C. Chiarcos, S. Hellmann and S. Nordhoff, Towards a linguistic linked open data cloud: The Open Linguistics Working Group, *Trait. Autom. des Langues* **52** (2011), 245–275.

[5] C. Bizer, T. Heath and T. Berners-Lee, Linked Data - The Story So Far, *International Journal on Semantic Web and Information Systems (IJSWIS)* **5**(3) (2009), 1–22. doi:10.4018/jswis.2009081901.

[6] C. Chiarcos, J. McCrae, P. Cimiano and C. Fellbaum, Towards open data for linguistics: Linguistic linked data, in: *New Trends of Research in Ontologies and Lexical Resources*, Springer, 2013, pp. 7–25. doi:10.1007/978-3-642-31782-8_2.

[7] J. Bosque-Gil, J. Gracia, E. Montiel-Ponsoda and A. Gómez-Pérez, Models to represent linguistic linked data, *Natural Language Engineering* **24**(6) (2018), 811–859. doi:10.1017/S1351324918000347.

[8] A.F. Khan, C. Chiarcos, T. Declerck, D. Gifu, E.G.-B. García, J. Gracia, M. Ionov, P. Labropoulou, F. Mambrini, J.P. McCrae, É. Pagé-Perron, M. Passarotti, R. Salvador and C.-O. Truică, When Linguistics Meets Web Technologies. Recent advances in Modelling Linguistic Linked Open Data, *Semantic Web* (2022). doi:10.3233/SW-222859.

[9] M.J. Page, J.E. McKenzie, P.M. Bossuyt, I. Boutron, T.C. Hoffmann, C.D. Mulrow, L. Shamseer, J.M. Tetzlaff, E.A. Akl, S.E. Brennan, R. Chou, J. Glanville, J.M. Grimshaw, A. Hróbjartsson, M.M. Lalu, T. Li, E.W. Loder, E. Mayo-Wilson, S. McDonald, L.A. McGuinness, L.A. Stewart, J. Thomas, A.C. Tricco, V.A. Welch, P. Whiting and D. Moher, The PRISMA 2020 statement: an updated guideline for reporting systematic reviews, *BMJ* **372** (2021). doi:10.1136/bmj.n71.

[10] M.D. Wilkinson, M. Dumontier, I.J. Aalbersberg, G. Appleton, M. Axton, A. Baak, N. Blomberg, J.-W. Boiten, L.B. da Silva Santos, P.E. Bourne, J. Bouwman, A.J. Brookes, T. Clark, M. Crosas, I. Dillo, O. Dumon, S. Edmunds, C.T. Evelo, R. Finkers, A. Gonzalez-Beltran, A.J.G. Gray, P. Groth, C. Goble, J.S. Grethe, J. Heringa, P.A.C. 't Hoen, R. Hooft, T. Kuhn, R. Kok, J. Kok, S.J. Lusher, M.E. Martone, A. Mons, A.L. Packer, B. Persson, P. Rocca-Serra, M. Roos, R. van Schaik, S.-A. Sansone, E. Schultes, T. Sengstag, T. Slater, G. Strawn, M.A. Swertz, M. Thompson, J. van der Lei, E. van Mulligen, J. Velterop, A. Waagmeester, P. Wittenburg, K. Wolstencroft, J. Zhao and B. Mons, The FAIR Guiding Principles for scientific data management and stewardship, *Scientific data* **3**(1) (2016), 1–9. doi:10.1038/sdata.2016.18.

[11] M. Ehrmann, F. Cecconi, D. Vannella, J.P. McCrae, P. Cimiano and R. Navigli, Representing Multilingual Data as Linked Data: the Case of BabelNet 2.0, in: *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, N. Calzolari, K. Choukri, T. Declerck, H. Loftsson, B. Maegaard, J. Mariani, A. Moreno, J. Odijk and S. Piperidis, eds, European Language Resources Association (ELRA), Reykjavik, Iceland, 2014, pp. 401–408.

[12] J. Gracia, E. Montiel-Ponsoda, P. Cimiano, A. Gómez-Pérez, P. Buitelaar and J. McCrae, Challenges for the multilingual Web of Data, *Journal of Web Semantics* **11** (2012), 63–71. doi:https://doi.org/10.1016/j.websem.2011.09.001.

[13] R. Cyganiak, D. Wood and M. Lanthaler, RDF 1.1 Concepts and Abstract Syntax, 2014. http://www.w3.org/TR/rdf11-concepts/.

[14] T. Berners-Lee, Linked Data, 2006-2010. https://www.w3.org/DesignIssues/LinkedData.html.

[15] P. Cimiano, C. Chiarcos, J.P. McCrae and J. Gracia, Linguistic Linked Data in Digital Humanities, in: *Linguistic Linked Data: Representation, Generation and Applications*, Springer International Publishing, Cham, 2020, pp. 229–262. doi:10.1007/978-3-030-30225-2_13.

[16] 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**(4) (2012), 701–719. doi:10.1007/s10579-012-9182-3.

[17] P. Cimiano, P. Haase, M. Herold, M. Mantel and P. Buitelaar, LexOnto: A Model for Ontology Lexicons for Ontology-based NLP, in: *Proceedings of the Workshop OntoLex - From Text to Knowledge: The Lexicon/Ontology Interface; held in conjunction with ISWC 2007*, 2007, pp. 1–12.

[18] P. Cimiano, C. Chiarcos, J.P. McCrae and J. Gracia, Linguistic Linked Open Data Cloud, in: *Linguistic Linked Data: Representation, Generation and Applications*, Springer International Publishing, Cham, 2020, pp. 29–41. ISBN 978-3-030-30225-2. doi:10.1007/978-3-030-30225-2_3.

[19] C. Chiarcos, S. Hellmann, S. Nordhoff, S. Moran, R. Littauer, J. Eckle-Kohler, I. Gurevych, S. Hartmann, M. Matuschek and C.M. Meyer, The Open Linguistics Working Group, in: *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, N. Calzolari, K. Choukri, T. Declerck, M.U. Doğan, B. Maegaard, J. Mariani, A. Moreno, J. Odijk and S. Piperidis, eds, European Language Resources Association (ELRA), Istanbul, Turkey, 2012, pp. 3603–3610.

[20] J.P. McCrae, C. Chiarcos, F. Bond, P. Cimiano, T. Declerck, G. de Melo, J. Gracia, S. Hellmann, B. Klimek, S. Moran, P. Osenova, A. Pareja-Lora and J. Pool, The Open Linguistics Working Group: Developing the Linguistic Linked Open Data Cloud, in: *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, 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), Portorož, Slovenia, 2016, pp. 2435–2441. https://aclanthology.org/L16-1386.

[21] B.C. Lust, M. Blume, A. Pareja-Lora and C. Chiarcos, Development of Linguistic Linked Open Data resources for collaborative data-intensive research in the language sciences: An introduction, A. Pareja-Lora, M. Blume, B.C. Lust and C. Chiarcos, eds, MIT Press, 2019, p. ix-xxi. ISBN 9780262536257.

[22] T. Declerck, Harmonizing Lexical Data for their Linking to Knowledge Objects in the Linked Data Framework, in: *Proceedings of Workshop on Lexical and Grammatical Resources for Language Processing*, J. Baptista, P. Bhattacharyya, C. Fellbaum, M. Forcada, C.-R. Huang, S. Koeva, C. Krstev and E. Laporte, eds, Association for Computational Linguistics and Dublin City University, Dublin, Ireland, 2014, pp. 18–23. doi:10.3115/v1/W14-5803. https://aclanthology.org/W14-5803.

[23] J.E. Labra Gayo, D. Kontokostas and S. Auer, Multilingual linked data patterns, *Semantic Web* **6**(4) (2015), 319–337. doi:10.3233/SW-140136.

[24] C. Chiarcos, S. Moran, P.N. Mendes, S. Nordhoff and R. Littauer, Building a Linked Open Data cloud of linguistic resources: Motivations and developments, *The People's Web Meets NLP* (2013). https://link.springer.com/chapter/10.1007/978-3-642-35085-6_12.

[25] C. Chiarcos, J. Mccrae, P. Osenova and C. Vertan, Linked Data in Linguistics 2014. Introduction and Overview, in: *3rd Workshop on Linked Data in Linguistics: Multilingual Knowledge Resources and Natural Language Processing*, C. Chiarcos, J. Mccrae, P. Osenova and C. Vertan, eds, 2014, p. vii–xv.

[26] S. Farrar and D.T. Langendoen, A linguistic ontology for the semantic web, *GLOT international* **7**(3) (2003), 97–100.

[27] P. Buitelaar, P. Cimiano, P. Haase and M. Sintek, Towards Linguistically Grounded Ontologies, in: *The Semantic Web: Research and Applications*, L. Aroyo, P. Traverso, F. Ciravegna, P. Cimiano, T. Heath, E. Hyvönen, R. Mizoguchi, E. Oren, M. Sabou and E. Simperl, eds, Springer Berlin Heidelberg, Berlin, Heidelberg, 2009, pp. 111–125.

[28] C. Chiarcos and M. Sukhareva, OLiA – Ontologies of Linguistic Annotation, *Semantic Web* **6** (2015), 379–386. doi:10.3233/SW-140167.

[29] T. Lesnikova, *NLP for interlinking multilingual LOD*, Proceedings of the ISWC Doctoral Consortium, HAL-Inria, Sydney, Australia, 2013, pp. 32–39, lesnikova2013b. https://hal.inria.fr/hal-00918496.

[30] T. Lesnikova, RDF Data Interlinking: Evaluation of Cross-lingual Methods, Theses, Université Grenoble Alpes, 2016. https://tel.archives-ouvertes.fr/tel-01366030.

[31] R. Lourdusamy and J. Florrence, Methods, approaches, principles, guidelines and applications on multilingual ontologies: a survey, *ICTACT Journal on Soft Computing* **7** (2016).

[32] V. Charles, H. Manguinhas, A. Isaac, N. Freire and S. Gordea, Designing a Multilingual Knowledge Graph as a Service for Cultural Heritage: Some Challenges and Solutions, in: *Proceedings of the 2018 International Conference on Dublin Core and Metadata Applications*, DCMI'18, Dublin Core Metadata Initiative, 2018, pp. 29–40–.

[33] N. Aggarwal, T. Polajnar and P. Buitelaar, Cross-Lingual Natural Language Querying over the Web of Data, in: *Natural Language Processing and Information Systems*, E. Métais, F. Meziane, M. Saraee, V. Sugumaran and S. Vadera, eds, Springer Berlin Heidelberg, Berlin, Heidelberg, 2013, pp. 152–163. ISBN 978-3-642-38824-8.

[34] S. Moran and C. Chiarcos, Linguistic Linked Open Data and Under-Resourced Languages: From Collection to Application, in: *Development of Linguistic Linked Open Data Resources for Collaborative Data-Intensive Research in the Language Sciences*, A. Pareja-Lora, M. Blume, B.C. Lust and C. Chiarcos, eds, MIT Press, 2019, pp. 39–68.

[35] C.-R. Huang, S.-K. Hsieh, L. Prévot, P.-Y. Hsiao and H.Y. Chang, Linking basic lexicon to shared ontology for endangered languages: a linked data approach toward Formosan languages, *Journal of Chinese Linguistics* **46** (2018).

[36] F. Gillis-Webber and S. Tittel, The Shortcomings of Language Tags for Linked Data When Modeling Lesser-Known Languages, in: *2nd Conference on Language, Data and Knowledge (LDK 2019)*, M. Eskevich, G. de Melo, C. Fäth, J.P. McCrae, P. Buitelaar, C. Chiarcos, B. Klimek and M. Dojchinovski, eds, OpenAccess Series in Informatics (OASIcs), Vol. 70, Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik, Dagstuhl, Germany, 2019, pp. 4:1–4:15. ISSN 2190-6807. ISBN 978-3-95977-105-4. doi:10.4230/OASIcs.LDK.2019.4.

[37] L. Pretorius, The Multilingual Semantic Web as Virtual Knowledge Commons: The Case of the Under-Resourced South African Languages, in: *Towards the Multilingual Semantic Web: Principles, Methods and Applications*, P. Buitelaar and P. Cimiano, eds, Springer Berlin Heidelberg, Berlin, Heidelberg, 2014, pp. 49–66. ISBN 978-3-662-43585-4. doi:10.1007/978-3-662-43585-4_4.

[38] Y. Li, Y. Yu and P. Fung, A Mandarin-English Code-Switching Corpus, in: *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, N. Calzolari, K. Choukri, T. Declerck, M.U. Doğan, B. Maegaard, J. Mariani, A. Moreno, J. Odijk and S. Piperidis, eds, European Language Resources Association (ELRA), Istanbul, Turkey, 2012, pp. 2515–2519.

[39] T. Schmidt and K. Wörner, *Multilingual corpora and multilingual corpus analysis*, Vol. 14, John Benjamins Publishing, 2012. doi:10.1075/hsm.14.

[40] M.-C. de Marneffe, C.D. Manning, J. Nivre and D. Zeman, Universal Dependencies, *Computational Linguistics* **47**(2) (2021), 255–308. doi:10.1162/coli_a_00402.

[41] J.P. McCrae, J. Bosque-Gil, J. Gracia, P. Buitelaar and P. Cimiano, The Ontolex-Lemon model: development and applications, in: *Electronic lexicography in the 21st century. 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., 2017, pp. 19–21. ISSN 2533-5626.

[42] P. Cimiano, C. Chiarcos, J.P. McCrae and J. Gracia, *Linguistic Linked Data - Representation, Generation and Applications*, Springer, 2020. ISBN 978-3-030-30224-5. doi:10.1007/978-3-030-30225-2.

[43] J. McCrae and T. Declerck, Linguistic Linked Open Data for All, *Proceedings of Language Technology 4 All* (2019), 13–15. doi:10.5281/zenodo.3607272. https://lt4all.elra.info/proceedings/lt4all2019/pdf/2019.lt4all-1.4.pdf.

[44] P. Buitelaar, T. Declerck, A. Frank, S. Racioppa, M. Kiesel, M. Sintek, R. Engel, M. Romanelli, D. Sonntag, B. Loos et al., Linginfo: Design and applications of a model for the integration of linguistic information in ontologies, in: *Proceedings of the OntoLex Workshop at LREC*, ELRA, 2006, pp. 28–32.

[45] M. Pazienza, A. Stellato and A. Turbati, Linguistic Watermark 3.0: an RDF framework and a software library for bridging language and ontologies in the Semantic Web, in: *5th Workshop on Semantic Web Applications and Perspectives, SWAP 2008*, Vol. 426, A. Aldo Gangemi, J. Keizer and H. Presutti Valentina ad Stoermer, eds, CEUR Workshop Proceedings, 2008.

[46] A. Oltramari and A. Stellato, Enriching ontologies with linguistic content: An evaluation framework, in: *Proceedings of OntoLex*, A. Oltramari, L. Prévot, C.-R. Huang and P. Vossen, eds, 2008.

[47] E. Monteil-Ponsoda, G. Aguado de Cea, A. Gómez-Pérez and W. Peters, Enriching ontologies with multilingual information, *Natural Language Engineering* **17**(3) (2011), 283–309. doi:10.1017/S1351324910000082.

[48] G. Francopoulo, M. George, N. Calzolari, M. Monachini, N. Bel, M. Pet and C. Soria, Lexical Markup Framework (LMF), 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), 2006.

[49] M. Kemps-Snijders, M. Windhouwer, P. Wittenburg and S.E. Wright, ISOcat: Corralling Data Categories in the Wild, in: *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, N. Calzolari, K. Choukri, B. Maegaard, J. Mariani, J. Odijk, S. Piperidis and D. Tapias, eds, European Language Resources Association (ELRA), Marrakech, Morocco, 2008.

[50] P. Cimiano, C. Chiarcos, J.P. McCrae and J. Gracia, Applying Linked Data Principles to Linking Multilingual Wordnets, in: *Linguistic Linked Data: Representation, Generation and Applications*, Springer International Publishing, Cham, 2020, pp. 215–228. doi:10.1007/978-3-030-30225-2_12.

[51] A. Fonseca, F. Sadat and F. Lareau, Lexfom: a lexical functions ontology model, in: *Proceedings of the 5th Workshop on Cognitive Aspects of the Lexicon (CogALex - V)*, M. Zock, A. Lenci and S. Evert, eds, The COLING 2016 Organizing Committee, Osaka, Japan, 2016, pp. 145–155.

[52] J.F. Sánchez-Rada and C.A. Iglesias, Onyx: A linked data approach to emotion representation, *Information Processing & Management* **52**(1) (2016), 99–114. doi:10.1016/j.ipm.2015.03.007.

[53] A. Miles and S. Bechhofer, SKOS simple knowledge organization system reference, *W3C recommendation* (2009).

[54] A. Gangemi, M. Alam, L. Asprino, V. Presutti and D.R. Recupero, Framester: A Wide Coverage Linguistic Linked Data Hub, in: *Knowledge Engineering and Knowledge Management*, E. Blomqvist, P. Ciancarini, F. Poggi and F. Vitali, eds, Springer International Publishing, Cham, 2016, pp. 239–254. ISBN 978-3-319-49004-5. doi:10.1007/978-3-319-49004-5_16.

[55] C.J. Fillmore et al., Frame semantics and the nature of language, in: *Annals of the New York Academy of Sciences: Conference on the origin and development of language and speech*, Vol. 280, New York, 1976, pp. 20–32. doi:10.1111/j.1749-6632.1976.tb25467.x.

[56] F. Abromeit, C. Fäth and L. Glaser, Annohub – Annotation Metadata for Linked Data Applications, in: *Proceedings of the 7th Workshop on Linked Data in Linguistics (LDL-2020)*, M. Ionov, J.P. McCrae, C. Chiarcos, T. Declerck, J. Bosque-Gil and J. Gracia, eds, European Language Resources Association (ELRA), Marseille, France, 2020, pp. 36–44. ISBN 979-10-95546-36-8.

[57] R. Loughnane, K. McCurdy, P. Kolb and S. Selent, Linked Data for Language-Learning Applications, in: *Proceedings of the 12th Workshop on Innovative Use of NLP for Building Educational Applications*, J. Tetreault, J. Burstein, C. Leacock and H. Yannakoudakis, eds, Association for Computational Linguistics, Copenhagen, Denmark, 2017, pp. 44–51. doi:10.18653/v1/W17-5005.

[58] V. Bryl, C. Bizer and H. Paulheim, Gathering Alternative Surface Forms for DBpedia Entities., in: *Proceedings of the Third NLP&DBpedia Workshop (NLP & DBpedia 2015)*, Vol. 1581, H. Paulheim, M. van Erp, A. Filipowska, P.N. Mendes and M. Brümmer, eds, CEUR Workshop Proceedings, 2015, pp. 13–24.

[59] R. Prokofyev, A. Tonon, M. Luggen, L. Vouilloz, D.E. Difallah and P. Cudré-Mauroux, SANAPHOR: Ontology-Based Coreference Resolution, in: *The Semantic Web - ISWC 2015*, M. Arenas, O. Corcho, E. Simperl, M. Strohmaier, M. d'Aquin, K. Srinivas, P. Groth, M. Dumontier, J. Heflin, K. Thirunarayan, K. Thirunarayan and S. Staab, eds, Springer International Publishing, Cham, 2015, pp. 458–473. doi:10.1007/978-3-319-25007-6_27.

[60] J. Plu, R. Prokofyev, A. Tonon, P. Cudré-Mauroux, D.E. Difallah, R. Troncy and G. Rizzo, Sanaphor++: Combining Deep Neural Networks with Semantics for Coreference Resolution, in: *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, N. Calzolari, K. Choukri, C. Cieri, T. Declerck, S. Goggi, K. Hasida, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, S. Piperidis and T. Tokunaga, eds, European Language Resources Association (ELRA), Miyazaki, Japan, 2018. https://aclanthology.org/L18-1063.

[61] B. Klimek, M. Ackermann, M. Brümmer and S. Hellmann, MMoOn Core-The Multilingual Morpheme Ontology, *Semantic Web Journal* **1**(5) (2021). doi:10.3233/SW-200412.

[62] B. Klimek, N. Arndt, S. Krause and T. Arndt, Creating Linked Data Morphological Language Resources with MMoOn - The Hebrew Morpheme Inventory (2016), 892–899.

[63] S. Racioppa and T. Declerck, Enriching Open Multilingual Wordnets with Morphological Features, in: *Proceedings of the Sixth Italian Conference on Computational Linguistics, Bari, Italy, November 13-15, 2019*, Vol. 2481, R. Bernardi, R. Navigli and G. Semeraro, eds, CEUR Workshop Proceedings, 2019.

[64] M. Sherif and A.N. Ngomo, Semantic Quran, *Semantic Web* **6** (2015), 339–345. doi:10.3233/SW-140137.

[65] C. Chiarcos and M. Ionov, Ligt: An LLOD-Native Vocabulary for Representing Interlinear Glossed Text as RDF, in: *2nd Conference on Language, Data and Knowledge (LDK 2019)*, M. Eskevich, G. de Melo, C. Fäth, J.P. McCrae, P. Buitelaar, C. Chiarcos, B. Klimek and M. Dojchinovski, eds, OpenAccess Series in Informatics (OASIcs), Vol. 70, Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik, Dagstuhl, Germany, 2019, pp. 3:1–3:15. ISSN 2190-6807. ISBN 978-3-95977-105-4. doi:10.4230/OASIcs.LDK.2019.3.

[66] M. Ionov, APiCS-Ligt: Towards Semantic Enrichment of Interlinear Glossed Text, in: *3rd Conference on Language, Data and Knowledge (LDK 2021)*, Schloss Dagstuhl-Leibniz-Zentrum für Informatik, 2021.

[67] S. Nordhoff, Modelling and Annotating Interlinear Glossed Text from 280 Different Endangered Languages as Linked Data with LIGT, in: *Proceedings of the 14th Linguistic Annotation Workshop*, S. Dipper and A. Zeldes, eds, Association for Computational Linguistics, Barcelona, Spain, 2020, pp. 93–104. https://aclanthology.org/2020.law-1.9.

[68] B. Klimek, J.P. McCrae, J. Bosque-Gil, M. Ionov, J.K. Tauber and C. Chiarcos, Challenges for the representation of morphology in ontology lexicons, in: *Electronic lexicography in the 21st century. Proceedings of the eLex 2019*, I. Kosem, T.Z. Kuhn, M. Correia, J.P. Ferreira, M. Jansen, I. Pereira, J. Kallas, M. Jakubíček, S. Krek and C. Tiberius, eds, Lexical Computing CZ, s.r.o., 2019, pp. 570–591.

[69] B. Klimek, Inducing the Cross-Disciplinary Usage of Morphological Language Data Through Semantic Modelling, PhD thesis, University of Basel, 2020.

[70] T. Declerck, M. Siegel and S. Racioppa, Using OntoLex-Lemon for Representing and Interlinking German Multiword Expressions in OdeNet and MMORPH, in: *Proceedings of the Joint Workshop on Multiword Expressions and WordNet (MWE-WN 2019)*, A. Savary, C.P. Escartín, F. Bond, J. Mitrović and V.B. Mititelu, eds, Association for Computational Linguistics, Florence, Italy, 2019, pp. 22–29. doi:10.18653/v1/W19-5104.

[71] C. Chiarcos, K. Donandt, M. Ionov, M. Rind-Pawlowski, H. Sargsian, J. Wichers Schreur, F. Abromeit and C. Fäth, Universal Morphologies for the Caucasus region, in: *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, N. Calzolari, K. Choukri, C. Cieri, T. Declerck, S. Goggi, K. Hasida, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, S. Piperidis and T. Tokunaga, eds, European Language Resources Association (ELRA), Miyazaki, Japan, 2018.

[72] M.C. Passarotti, F.M. Cecchini, G. Franzini, E. Litta, F. Mambrini and P. Ruffolo, The LiLa Knowledge Base of Linguistic Resources and NLP Tools for Latin., in: *LDK-PS 2019*, Vol. 2402, T. Decklerck and J. McCrae, eds, CEUR Workshop Proceedings, 2019, pp. 6–11.

[73] A. Pareja-Lora, OntoLingAnnot's Ontologies: Facilitating Interoperable Linguistic Annotations (Up to the Pragmatic Level), in: *Linked Data in Linguistics*, C. Chiarcos, S. Nordhoff and S. Hellmann, eds, Springer, 2012, pp. 117–127. doi:10.1007/978-3-642-28249-2_12.
[74] C. Chiarcos, Towards interoperable discourse annotation. Discourse features in the Ontologies of Linguistic Annotation, in: *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, N. Calzolari, K. Choukri, T. Declerck, H. Loftsson, B. Maegaard, J. Mariani, A. Moreno, J. Odijk and S. Piperidis, eds, European Language Resources Association (ELRA), Reykjavik, Iceland, 2014, pp. 4569–4577. http://www.lrec-conf.org/proceedings/lrec2014/pdf/893_Paper.pdf.
[75] C. Chiarcos, An ontology of linguistic annotations., in: *LDV Forum*, Vol. 23, Citeseer, 2008, pp. 1–16.
[76] M. Bärenfänger, M. Hilbert, H. Lobin and H. Lüngen, OWL ontologies as a resource for discourse parsing, *LDV-Forum* **1**(23) (2008), 17–26.
[77] D. Goecke, H. Lüngen, F. Sasaki, A. Witt and S. Farrar, GOLD and discourse: Domain-and community-specific extensions, in: *Proceedings of the E-MELD Workshop on Morphosyntactic Annotation and Terminology: Linguistic Ontologies and Data Categories for Language Resources*, E-MELD, 2005.
[78] H. Bunt and R. Prasad, ISO DR-Core (ISO 24617-8): Core concepts for the annotation of discourse relations, in: *Proceedings 12th Joint ACL-ISO Workshop on Interoperable Semantic Annotation (ISA-12)*, 2016, pp. 45–54.
[79] J. Hoek, J. Evers-Vermeul and T.J. Sanders, Using the cognitive approach to coherence relations for discourse annotation, *Dialogue & Discourse* **10**(2) (2019), 1–33. doi:10.5087/dad.2019.201.
[80] G. Valūnaitė Oleškevičienė, C. Liebeskind, D. Trajanov, P. Silvano, C. Chiarcos and M. Damova, Speaker Attitudes Detection through Discourse Markers Analysis, in: *Proceedings of Workshop on Deep learning and Neural Approaches for Lingusitic Data*, R. Garabík, ed., NexusLinguarum, 2021, pp. 8–12.
[81] M. Mladenović and J. Mitrović, Ontology of rhetorical figures for Serbian, in: *Text, Speech, and Dialogue. TSD 2013*, Vol. 8082, I. Habernal and V. Matoušek, eds, Springer, 2013, pp. 386–393. doi:10.1007/978-3-642-40585-3_49.
[82] T. Nurmikko-Fuller, Assessing the Suitability of Existing owl Ontologies for the Representation of Narrative Structures in Sumerian Literature, *ISAW Papers* **7**(18) (2014).
[83] F. Branch, T. Arias, J. Kennah, R. Phillips, T. Windleharth and J.H. Lee, Representing transmedia fictional worlds through ontology, *Journal of the Association for Information Science and Technology* **68**(12) (2017), 2771–2782. doi:10.1002/asi.23886.
[84] J. Mitrović, C. O'Reilly, M. Mladenović and S. Handschuh, Ontological representations of rhetorical figures for argument mining, *Argument & Computation* **8**(3) (2017), 267–287. doi:10.3233/AAC-170027.
[85] H. Bermúdez-Sabel, M.L. Díez Platas, S. Ros and E. González-Blanco, Towards a common model for European Poetry: Challenges and solutions, *Digital Scholarship in the Humanities* (2021). doi:10.1093/llc/fqab106.
[86] M. Damova, D. Dannélls, R. Enache, M. Mateva and A. Ranta, Multilingual natural language interaction with semantic web knowledge bases and linked open data, in: *Towards the Multilingual Semantic Web*, P. Buitelaar and P. Cimiano, eds, Springer, 2014, pp. 211–226. doi:10.1007/978-3-662-43585-4_13.
[87] S. Hakimov, S. Jebbara and P. Cimiano, AMUSE: multilingual semantic parsing for question answering over linked data, in: *International Semantic Web Conference*, C. d'Amato, M. Fernandez, V. Tamma, F. Lecue, P. Cudré-Mauroux, J. Sequeda, C. Lange and J. Heflin, eds, Springer, 2017, pp. 329–346. doi:10.1007/978-3-319-68288-4_20.
[88] L.-A. Kaffee, K.M. Endris, E. Simperl and M.-E. Vidal, Ranking Knowledge Graphs By Capturing Knowledge about Languages and Labels, in: *Proceedings of the 10th International Conference on Knowledge Capture*, K-CAP '19, Association for Computing Machinery, New York, NY, USA, 2019, pp. 21–28–. ISBN 9781450370080. doi:10.1145/3360901.3364443.
[89] G. Wilcock, Talking OWLs: Towards an ontology verbalizer, 2003, pp. 109–112. https://gate.ac.uk/conferences/iswc2003/proceedings/.
[90] G. Wilcock, An OWL Ontology for HPSG, in: *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, S. Ananiadou, ed., Association for Computational Linguistics, Prague, Czech Republic, 2007, pp. 169–172. https://aclanthology.org/P07-2043.
[91] D.A. Cojocaru and S. Trausan-Matu, Text Generation Starting from an Ontology., in: *12th Romanian Human-Computer Interaction Conference*, RoCHI, 2015, pp. 55–60. ISSN 2501-9422.
[92] R.A. Harris, C. Di Marco, A.R. Mehlenbacher, R. Clapperton, I. Choi, I. Li, S. Ruan and C. O'Reilly, A cognitive ontology of rhetorical figures, in: *Proceedings of AISB Annual Convention 2017, Symposium on Cognition and Ontologies (CAOS)*, J. Bryson, M.D. Vos, and J. Padget, eds, Society for the Study of Artificial Intelligence & Simulation of Behaviour, 2017, pp. 228–235. ISBN 978-1-908187-29-1.
[93] R.A. Harris, C. Di Marco, S. Ruan and C. O'Reilly, An annotation scheme for rhetorical figures, *Argument & Computation* **9**(2) (2018), 155–175. doi:10.3233/AAC-180037.
[94] Y. Wang, R.A. Harris and D.M. Berry, An Ontology for Ploke: Rhetorical Figures of Lexical Repetitions, in: *Proceedings of the Joint Ontology Workshops 2021: Episode VII The Bolzano Summer of Knowledge*, Vol. 2969, E.M. Sanfilippo, O. Kutz, N. Troquard, T. Hahmann, C. Masolo, R. Hoehndorf and R. Vita, eds, CEUR Workshop Proceedings, 2021. ISSN 1613-0073.
[95] C. Chiarcos and M. Ionov, Linking Discourse Marker Inventories, 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. 40:1–40:15. ISSN 2190-6807. ISBN 978-3-95977-199-3. doi:10.4230/OASIcs.LDK.2021.40.
[96] J. Bosque-Gil, J. Gracia and A. Gómez-Pérez, Linked data in lexicography, *Kernerman DICTIONARY News* (2016), 19–24.
[97] B. Klimek and M. Brümmer, Enhancing lexicography with semantic language databases, *Kernerman DICTIONARY News* **23** (2015), 5–10. https://www.kdictionaries.com/kdn/kdn23_2015.pdf.
[98] J. Gracia, M. Villegas, A. Gomez-Perez and N. Bel, The apertium bilingual dictionaries on the web of data, *Semantic Web* **9**(2) (2018), 231–240. doi:10.3233/SW-170258.

[99] J. Bosque-Gil, J. Gracia, E. Montiel-Ponsoda and G. Aguado-de-Cea, Modelling Multilingual Lexicographic Resources for the Web of Data: the K Dictionaries case, in: *Proceedings of GLOBALEX'16 workshop at LREC'15, Portoroz, Slovenia*, I. Kernerman, I. Kosem, S. Krek and L. Trap-Jensen, eds, European Language Resources Association (ELRA), 2016. ISBN 978-2-9517408-9-1.

[100] F. Khan, J.E. Díaz-Vera and M. Monachini, Representing Polysemy and Diachronic Lexico-Semantic Data on the Semantic Web, in: *Proceedings of the Second International Workshop on Semantic Web for Scientific Heritage co-located with 13th Extended Semantic Web Conference (ESWC 2016)*, Vol. 1595, I. Draelants, C.F. Zucker, A. Monnin and A. Zucker, eds, CEUR Workshop Proceedings, Heraklion, 2016, pp. 37–46.

[101] T. Declerck and E. Wandl-Vogt, Cross-linking Austrian dialectal Dictionaries through formalized Meanings, in: *Proceedings of the XVI EURALEX International Congress: The User in Focus*, A. Abel, C. Vettori and N. Ralli, eds, 2014, pp. 329–343.

[102] F. Abromeit, C. Chiarcos, C. Fäth and M. Ionov, Linking the Tower of Babel: modelling a massive set of etymological dictionaries as RDF, in: *Proceedings of the 5th Workshop on Linked Data in Linguistics (LDL-2016): Managing, Building and Using Linked Language Resources*, J.P. McCrae, C. Chiarcos, E.M. Ponsoda, T. Declerck, P. Osenova and S. Hellmann, eds, 2016, pp. 11–19.

[103] J. Gracia, Multilingual dictionaries and the Web of Data, *Kernerman DICTIONARY News* (2015), 1–4.

[104] J. Gracia, C. Fäth, M. Hartung, M. Ionov, J. Bosque-Gil, S. Veríssimo, C. Chiarcos and M. Orlikowski, Leveraging Linguistic Linked Data for Cross-Lingual Model Transfer in the Pharmaceutical Domain, in: *The Semantic Web – ISWC 2020*, Vol. 12507, J.Z. Pan, V. Tamma, C. d'Amato, K. Janowicz, B. Fu, A. Polleres, O. Seneviratne and L. Kagal, eds, Springer, 2020, pp. 499–514. doi:10.1007/978-3-030-62466-8_31.

[105] J. Bosque-Gil, J. Gracia and E. Montiel-Ponsoda, Towards a Module for Lexicography in OntoLex, in: *LDK Workshops 2017: OntoLex, TIAD and Challenges for Wordnets*, Vol. 1899, J.P. McCrae, F. Bond, P. Buitelaar, P. Cimiano, T. Declerck, J. Gracia, I. Kernerman, E.M. Ponsoda, N. Ordan, and M. Piasecki, eds, CEUR Workshop Proceedings, 2017, pp. 74–84.

[106] J. Bosque-Gil, D. Lonke, I. Kernerman and J. Gracia, Validating the ontolex-lemon lexicography module with K dictionaries' multilingual data, in: *Electronic lexicography in the 21st century. Proceedings of the eLex 2019 conference*, I. Kosem, T.Z. Kuhn, M. Correia, J.P. Ferreira, M. Jansen, I. Pereira, J. Kallas, M. Jakubíček, S. Krek and C. Tiberius, eds, Lexical Computing CZ, s.r.o., 2019, pp. 726–746. ISSN 2533-5626.

[107] F. Mambrini, E. Litta, M. Passarotti and P. Ruffolo, Linking the Lewis & Short Dictionary to the LiLa Knowledge Base of Interoperable Linguistic Resources for Latin, in: *CLiC-it 2021 - Proceedings of the Eighth Italian Conference on Computational Linguistics*, Vol. 3033, E. Fersini, M. Passarotti and V. Patti, eds, CEUR Workshop Proceedings, 2021. doi:10.5281/ZENODO.5773783.

[108] T. Declerck, J.P. McCrae, M. Hartung, J. Gracia, C. Chiarcos, E. Montiel-Ponsoda, P. Cimiano, A. Revenko, R. Sauri, D. Lee et al., Recent developments for the linguistic linked open data infrastructure, in: *Proceedings of the 12th Language Resources and Evaluation Conference*, N. Calzolari, F. Béchet, P. Blache, K. Choukri, C. Cieri, T. Declerck, S. Goggi, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk and S. Piperidis, eds, European Language Resources Association (ELRA), 2020, pp. 5660–5667. ISBN 979-10-95546-34-4.

[109] J.P. McCrae, C. Tiberius, A.F. Khan, I. Kernerman, T. Declerck, S. Krek, M. Monachini and S. Ahmadi, The ELEXIS interface for interoperable lexical resources, in: *Electronic lexicography in the 21st century. Proceedings of the eLex 2019 conference*, I. Kosem, T.Z. Kuhn, M. Correia, J.P. Ferreira, M. Jansen, I. Pereira, J. Kallas, M. Jakubíček, S. Krek and C. Tiberius, eds, Lexical Computing CZ, s.r.o., 2019, pp. 642–659.

[110] T. Declerck, C. Tiberius and E. Wandl-Vogt, Encoding lexicographic data in lemon: Lessons learned, in: *Proceedings of the LDK workshops: OntoLex, TIAD and Challenges for Wordnets*, Vol. 1899, J.P. McCrae, F. Bond, P. Buitelaar, P. Cimiano, T. Declerck, J. Gracia, I. Kernerman, E.M. Ponsoda, N.O. 6 and M. Piasecki, eds, CEUR Workshop Proceedings, 2017.

[111] F. Khan, Towards the Representation of Etymological and Diachronic Lexical Data on the Semantic Web, in: *Proceedings of the 6th Workshop on Linked Data in Linguistics LDL*, J.P. McCrae, C. Chiarcos, T. Declerck, J. Gracia and B. Klimek, eds, European Language Resources Association (ELRA), 2018.

[112] G. de Melo, Etymological Wordnet: Tracing The History of Words, in: *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, N. Calzolari, K. Choukri, T. Declerck, H. Loftsson, B. Maegaard, J. Mariani, A. Moreno, J. Odijk and S. Piperidis, eds, European Language Resources Association (ELRA), Reykjavik, Iceland, 2014, pp. 1148–1154. https://aclanthology.org/L14-1.

[113] E. Pantaleo, V.W. Anelli, T. Di Noia and G. Serasset, Etytree: A Graphical and Interactive Etymology Dictionary Based on Wiktionary, in: *WWW '17 Companion: Proceedings of the 26th International Conference on World Wide Web Companion*, International World Wide Web Conferences Steering Committee, Perth, Australia, 2017, pp. 1635–1640. doi:10.1145/3041021.3053365.

[114] G. Sérasset, DBnary: Wiktionary as a Lemon-Based Multilingual Lexical Resource in RDF, *Semantic Web* **6**(4) (2015), 355–361. doi:10.3233/SW-140147.

[115] C. Chiarcos and M. Sukhareva, Linking etymological databases. A case study in Germanic, in: *3rd Workshop on Linked Data in Linguistics: Multilingual Knowledge Resources and Natural Language Processing*, C. Chiarcos, J.P. McCrae, P. Osenova and C. Vertan, eds, European Language Resources Association (ELRA), 2014, pp. 41–49.

[116] A. Bellandi, E. Giovannetti and A. Weingart, Multilingual and multiword phenomena in a lemon old occitan medico-botanical lexicon, *Information* **9**(3) (2018), 52. doi:10.3390/info9030052.

[117] R. Gennari and T. Di Mascio, An Ontology for a Web Dictionary of Italian Sign Language, in: *WEBIST 2007 - Proceedings of the Third International Conference on Web Information Systems and Technologies*, Vol. WIA, J. Filipe, J. Cordeiro, B. Encarnação and V. Pedrosa, eds, INSTICC Press, 2007, pp. 206–213.

[118] T. Homburg, PaleoCodage—Enhancing machine-readable cuneiform descriptions using a machine-readable paleographic encoding, *Digital Scholarship in the Humanities* **36**(Supplement_2) (2021), ii127–ii154.

[119] S. Tittel and F. Gillis-Webber, Identification of Languages in Linked Data: A Diachronic-Diatopic Case Study of French, in: *Electronic lexicography in the 21st century. Proceedings of the eLex 2019 conference*, I. Kosem, T.Z. Kuhn, M. Correia, J.P. Ferreira, M. Jansen, I. Pereira, J. Kallas, M. Jakubíček, S. Krek and C. Tiberius, eds, Lexical Computing CZ, s.r.o., 2019, pp. 547–569. https://elex.link/elex2019/proceedings-download/.

[120] S. Moran, Using Linked Data to Create a Typological Knowledge Base, in: *Linked Data in Linguistics - Representing and Connecting Language Data and Language Metadata*, C. Chiarcos, S. Nordhoff and S. Hellmann, eds, Springer, 2012, pp. 129–138. doi:10.1007/978-3-642-28249-2_13.

[121] D. Vila-Suero, A. Gómez-Pérez, E. Montiel-Ponsoda, J. Gracia and G. Aguado-de-Cea, Publishing linked data on the web: The multilingual dimension, in: *Towards the Multilingual Semantic Web*, P. Buitelaar and P. Cimiano, eds, Springer, 2014, pp. 101–117. doi:10.1007/978-3-662-43585-4_7.

[122] B. Villazón-Terrazas, L.M. Vilches-Blázquez, O. Corcho and A. Gómez-Pérez, Methodological guidelines for publishing government linked data, in: *Linking government data*, D. Wood, ed., Springer, 2011, pp. 27–49. doi:10.1007/978-1-4614-1767-5_2.

[123] J. Gracia, E. Montiel-Ponsoda, D. Vila-Suero and G. Aguado-de-Cea, Enabling Language Resources to Expose Translations as Linked Data on the Web, in: *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, N. Calzolari, K. Choukri, T. Declerck, H. Loftsson, B. Maegaard, J. Mariani, A. Moreno, J. Odijk and S. Piperidis, eds, European Language Resources Association (ELRA), Reykjavik, Iceland, 2014, pp. 409–413. https://aclanthology.org/L14-1.

[124] E. Montiel-Ponsoda, J. Gracia, G. Aguado-de-Cea and A. Gómez-Pérez, Representing Translations on the Semantic Web, in: *Proceedings of the 2nd International Workshop on the Multilingual Semantic Web (MSW 2011)*, Vol. 1755, E. Montiel-Ponsoda, J. McCrae, P. Buitelaar and P. Cimiano, eds, CEUR Workshop Proceedings, 2011, pp. 25–37–. ISSN 1613-0073.

[125] Z. Fang, H. Wang, J. Gracia, J. Bosque-Gil and T. Ruan, Zhishi. lemon: On publishing Zhishi. me as linguistic linked open data, in: *International Semantic Web Conference*, P. Groth, E. Simperl, A. Gray, M. Sabou, M. Krötzsch, F. Lecue, F. Flöck and Y. Gil, eds, Springer, Cham, 2016, pp. 47–55. doi:10.1007/978-3-319-46547-0_6.

[126] D. Moussallem, M.A. Sherif, D. Esteves, M. Zampieri and A.N. Ngomo, LIDIOMS: A Multilingual Linked Idioms Data Set, *CoRR* **abs/1802.08148** (2018). http://arxiv.org/abs/1802.08148.

[127] P. León-Araúz and P. Faber, Context and Terminology in the Multilingual Semantic Web, in: *Towards the Multilingual Semantic Web*, Springer Berlin Heidelberg, 2014, pp. 31–47. doi:10.1007/978-3-662-43585-4_3.

[128] C. Federmann, D. Gromann, T. Declerck, S. Hunsicker, H.-U. Krieger and G. Budin, Multilingual Terminology Acquisition for Ontology-based Information Extraction, in: *Proceedings of the 10th Terminology and Knowledge Engineering Conference*, TKE, 2012, pp. 166–175.

[129] M. Arcan and P. Buitelaar, MONNET: Multilingual Ontologies for Networked Knowledge, in: *Proceedings of Machine Translation Summit XIV: European projects*, A. Way, ed., Nice, France, 2013. https://aclanthology.org/2013.mtsummit-european.13.

[130] H.-U. Krieger and T. Declerck, TMO — The Federated Ontology of the TrendMiner Project, in: *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, N. Calzolari, K. Choukri, T. Declerck, H. Loftsson, B. Maegaard, J. Mariani, A. Moreno, J. Odijk and S. Piperidis, eds, European Language Resources Association (ELRA), Reykjavik, Iceland, 2014, pp. 4164–4171. http://www.lrec-conf.org/proceedings/lrec2014/pdf/115_Paper.pdf.

[131] D. Lewis, Position Paper: Interoperability Challenges for Linguistic Linked Data, in: *Proceedings of the W3C Workshop on Open Data on the Web*, W3C, 2013.

[132] J. Bosque-Gil, J. Gracia, G. Aguado-de-Cea and E. Montiel-Ponsoda, Applying the Ontolex model to a multilingual terminological resource, in: *Proc. of 12th Extended Semantic Web Conference (ESWC 2015) Satellite Events, Portorož, Slovenia*, Lecture Notes in Computer Science, Vol. 9341, Springer, 2015, pp. 283–294. ISBN 9783319256382. doi:10.1007/978-3-319-25639-9_43.

[133] M.P. di Buono, P. Cimiano, M.F. Elahi and F. Grimm, Terme-à-LLOD: Simplifying the Conversion and Hosting of Terminological Resources as Linked Data, in: *Proceedings of the 7th Workshop on Linked Data in Linguistics (LDL-2020)*, M. Ionov, J.P. McCrae, C. Chiarcos, T. Declerck, J. Bosque-Gil and J. Gracia, eds, European Language Resources Association, Marseille, France, 2020, pp. 28–35. ISBN 979-10-95546-36-8. https://aclanthology.org/2020.ldl-1.5.

[134] L. Wachowiak, C. Lang, B. Heinisch and D. Gromann, Towards Learning Terminological Concept Systems from Multilingual Natural Language Text, 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. 22:1–22:18. ISSN 2190-6807. ISBN 978-3-95977-199-3. doi:10.4230/OASIcs.LDK.2021.22. https://drops.dagstuhl.de/opus/volltexte/2021/14558.

[135] G. Speranza, C. Carlino and S. Ahmadi, Creating a Multilingual Terminological Resource using Linked Data: the case of Archaeological Domain in the Italian language., in: *CLiC-it 2019 Italian Conference on Computational Linguistics*, Vol. 2481, R. Bernardi, R. Navigli and G. Semeraro, eds, CEUR Workshop Proceedings, 2019. ISSN 1613–0073.

[136] M. Blume, A. Pareja-Lora, S. Flynn, C. Foley, T. Caldwell, J. Reidy, J. Masci and B. Lust, 9: Enabling New Collaboration and Research Capabilities in Language Sciences: Management of Language Acquisition Data and Metadata with the Data Transcription and Analysis Tool, A. Pareja-Lora, M. Blume, B.C. Lust and C. Chiarcos, eds, MIT Press, 2019. doi:10.7551/mitpress/10990.003.0011.

[137] L. Lezcano, S. Sánchez-Alonso and A.J. Roa-Valverde, A survey on the exchange of linguistic resources, *Program: electronic library and information systems* **47** (2013).

[138] N. Ide and K. Suderman, GrAF: A Graph-based Format for Linguistic Annotations, in: *Proceedings of the Linguistic Annotation Workshop*, B. Boguraev, N. Ide, A. Meyers, S. Nariyama, M. Stede, J. Wiebe and G. Wilcock, eds, Association for Computational Linguistics, Prague, Czech Republic, 2007, pp. 1–8. https://aclanthology.org/W07-1501.

[139] T.E.I. Consortium, TEI P5: Guidelines for electronic text encoding and interchange, 2008. http://www.tei-c.org/release/doc/tei-p5-doc/en/html/SG.html.
[140] M. Jewell, Semantic screenplays: Preparing TEI for linked data, *Digital Humanities* **2010** (2010), 2010.
[141] C. Chiarcos and T. Erjavec, OWL/DL formalization of the MULTEXT-East morphosyntactic specifications, in: *Proceedings of the 5th Linguistic Annotation Workshop*, 2011, pp. 11–20.
[142] A. Witt, U. Heid, F. Sasaki and G. Sérasset, Multilingual language resources and interoperability, *Language Resources and Evaluation* **43**(1) (2009), 1–14. doi:10.1007/s10579-009-9088-x.
[143] P. Buitelaar and P. Cimiano, *Towards the multilingual semantic web*, Springer, 2014. doi:10.1007/978-3-662-43585-4.
[144] R. Sanderson, P. Ciccarese and B. Young, Web annotation data model, *W3C recommendation* **23** (2017).
[145] S. Hellmann, J. Lehmann, S. Auer and M. Brümmer, Integrating NLP Using Linked Data, in: *The Semantic Web - ISWC 2013*, H. Alani, L. Kagal, A. Fokoue, P. Groth, C. Biemann, J.X. Parreira, L. Aroyo, N. Noy, C. Welty and K. Janowicz, eds, Springer, 2013. doi:10.1007/978-3-642-41338-4_7.
[146] A. Burchardt, S. Padó, D. Spohr, A. Frank and U. Heid, Formalising Multi-layer Corpora in OWL DL - Lexicon Modelling, Querying and Consistency Control, in: *Proceedings of the Third International Joint Conference on Natural Language Processing: Volume-I*, 2008. https://aclanthology.org/I08-1051.
[147] U. Czeitschner, T. Declerck and C. Resch, Porting Elements of the Austrian Baroque Corpus onto the Linguistic Linked Open Data Format, in: *Proceedings of the Joint Workshop on NLP&LOD and SWAIE: Semantic Web, Linked Open Data and Information Extraction*, D. Maynard, M. van Erp, B. Davis, P. Osenova, K. Simov, G. Georgiev and P. Nakov, eds, INCOMA Ltd. Shoumen, BULGARIA, Hissar, Bulgaria, 2013, pp. 12–15. https://aclanthology.org/W13-5204.
[148] V. Dimitrova and H. Renner-Westermann, Das Linguistik-Portal: Übergang von einer Virtuellen Fachbibliothek zu einem Fachinformationsdienst, *Bibliotheksdienst* **52**(3–4) (2018), 278–289. doi:10.1515/bd-2018-0033.
[149] C. Chiarcos, M. Ionov, M. Rind-Pawlowski, C. Fäth, J.W. Schreur and I. Nevskaya, LLODifying linguistic glosses, in: *International Conference on Language, Data and Knowledge*, J. Gracia, F. Bond, J.P. McCrae, P. Buitelaar, C. Chiarcos and S. Hellmann, eds, Springer, 2017, pp. 89–103. doi:10.1007/978-3-319-59888-8_7.
[150] D. Mukhamedshin, O. Nevzorova and A. Kirillovich, Using FLOSS for Storing, Processing and Linking Corpus Data, in: *Open Source Systems. OSS 2020. IFIP Advances in Information and Communication Technology*, V. Ivanov, A. Kruglov, S. Masyagin, A. Sillitti and G. Succi, eds, Springer, 2020, pp. 177–182. doi:10.1007/978-3-030-47240-5_17.
[151] C. Chiarcos, Interoperability of corpora and annotations, in: *Linked Data in Linguistics*, C. Chiarcos, S. Nordhoff and S. Hellmann, eds, Springer, 2012, pp. 161–179. doi:0.1007/978-3-642-28249-2_16.
[152] C. Pollin, G. Schneider, K. Gerhalter and M. Hummel, Semantic Annotation in the Project "Open Access Database 'Adjective-Adverb Interfaces' in Romance", in: *annDH 2018 Annotation in Digital Humanities*, Vol. 2155, S. Kübler and H. Zinsmeister, eds, CEUR Workshop Proceedings, 2018, pp. 41–46. ISSN 1613–0073.
[153] C. Chiarcos, POWLA: Modeling Linguistic Corpora in OWL/DL, in: *The Semantic Web: Research and Applications - 9th Extended Semantic Web Conference, ESWC 2012, Heraklion, Crete, Greece, May 27-31, 2012. Proceedings*, E. Simperl, P. Cimiano, A. Polleres, Ó. Corcho and V. Presutti, eds, Lecture Notes in Computer Science, Vol. 7295, Springer, 2012, pp. 225–239. doi:10.1007/978-3-642-30284-8_22.
[154] E.W. De Luca, Extending the Linked Data Cloud with Multilingual Lexical Linked Data, *KNOWLEDGE ORGANIZATION* **40**(5) (2013), 320–331. doi:10.5771/0943-7444-2013-5-320.
[155] J.P. McCrae and P. Buitelaar, Linking datasets using semantic textual similarity, *Cybernetics and information technologies* **18**(1) (2018), 109–123.
[156] F. Gillis-Webber, The construction of a linguistic linked data framework for bilingual lexicographic resources, PhD thesis, University of Cape Town, 2018.
[157] C. Caracciolo, A. Stellato, S. Rajbahndari, A. Morshed, G. Johannsen, J. Keizer and Y. Jaques, Thesaurus Maintenance, Alignment and Publication as Linked Data. The AGROVOC Use Case, *International Journal of Metadata, Semantics and Ontologies* **7**(1) (2012), 65. doi:10.1504/IJMSO.2012.048511.
[158] R. Albertoni, M.D. Martino, P. Podestà, A. Abecker, R. Wössner and K. Schnitter, LusTRE: a framework of linked environmental thesauri for metadata management, *Earth Science Informatics* **11**(4) (2018), 525–544. doi:10.1007/s12145-018-0344-8.
[159] J. Lehmann, R. Isele, M. Jakob, A. Jentzsch, D. Kontokostas, P.N. Mendes, S. Hellmann, M. Morsey, P. van Kleef, S. Auer and et al., DBpedia – A large-scale, multilingual knowledge base extracted from Wikipedia, *Semantic Web* **6**(2) (2015), 167–195. doi:10.3233/SW-140134.
[160] S. Hellmann, J. Brekle and S. Auer, Leveraging the Crowdsourcing of Lexical Resources for Bootstrapping a Linguistic Data Cloud, in: *Semantic Technology*, H. Takeda, Y. Qu, R. Mizoguchi and Y. Kitamura, eds, Springer Berlin Heidelberg, Berlin, Heidelberg, 2013, pp. 191–206. ISBN 978-3-642-37996-3.
[161] T. Flati, Learning of a multilingual bitaxonomy of Wikipedia and its application to semantic predicates (2015).
[162] R. Steinberger, M. Ebrahim, A. Poulis, M. Carrasco-Benitez, P. Schlüter, M. Przybyszewski and S. Gilbro, An overview of the European Union's highly multilingual parallel corpora, *Language Resources and Evaluation* **48**(4) (2014), 679–707. doi:10.1007/s10579-014-9277-0.
[163] R. Steinberger, B. Pouliquen, A. Widiger, C. Ignat, T. Erjavec, D. Tufiş and D. Varga, The JRC-Acquis: A Multilingual Aligned Parallel Corpus with 20+ Languages, in: *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*, European Language Resources Association (ELRA), 2006. http://www.lrec-conf.org/proceedings/lrec2006/pdf/340_pdf.pdf.

[164] N. Hajlaoui, D. Kolovratnik, J. Väyrynen, R. Steinberger and D. Varga, DCEP - Digital Corpus of the European Parliament, in: *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, N. Calzolari, K. Choukri, T. Declerck, H. Loftsson, B. Maegaard, J. Mariani, A. Moreno, J. Odijk and S. Piperidis, eds, European Language Resources Association (ELRA), 2014. http://www.lrec-conf.org/proceedings/lrec2014/pdf/943_Paper.pdf.

[165] R. Steinberger, A. Eisele, S. Klocek, S. Pilos and P. Schlüter, DGT-TM: A freely available Translation Memory in 22 languages, in: *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, N. Calzolari, K. Choukri, T. Declerck, M.U. Doğan, B. Maegaard, J. Mariani, A. Moreno, J. Odijk, N. Piperidis SteliosCalzolari, K. Choukri, T. Declerck, M.U. Doğan, B. Maegaard, J. Mariani, A. Moreno, J. Odijk and S. Piperidis, eds, European Language Resources Association (ELRA), 2012, pp. 454–459. http://www.lrec-conf.org/proceedings/lrec2012/pdf/814_Paper.pdf.

[166] R. Steinberger, M. Ebrahim and M. Turchi, JRC Eurovoc Indexer JEX - A freely available multi-label categorisation tool, in: *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, N. Calzolari, K. Choukri, T. Declerck, M.U. Doğan, B. Maegaard, J. Mariani, A. Moreno, J. Odijk, N. Piperidis SteliosCalzolari, K. Choukri, T. Declerck, M.U. Doğan, B. Maegaard, J. Mariani, A. Moreno, J. Odijk and S. Piperidis, eds, European Language Resources Association (ELRA), 2012, pp. 798–805. http://www.lrec-conf.org/proceedings/lrec2012/pdf/875_Paper.pdf.

[167] R. Steinberger, Multilingual and Cross-Lingual News Analysis in the Europe Media Monitor (EMM) (Extended Abstract), in: *The 6th Information Retrieval Facility Conference (IRFC'2013)*, Springer Berlin Heidelberg, 2013, pp. 1–4. doi:10.1007/978-3-642-41057-4_1.

[168] P.L. Araúz, P.J.M. Redondo and P. Faber, Integrating Environment into the Linked Data Cloud., in: *EnviroInfo*, 2011, pp. 370–379.

[169] G.M. Di Nunzio and S. Rabanus, Research on geolinguistic linked data: The test case of Cimbrian varieties, *Tosques F (ed.)* **20** (2014), 1–8.

[170] L.A. Díez, B. Pérez-León, M. Martínez-González and D.-J.V. Blanco, Propuesta de representación del tesauro Eurovoc en SKOS para su integración en sistemas de información jurídica, *Scire: representación y organización del conocimiento* (2010), 47–51.

[171] P. Martín-Chozas, E. Montiel-Ponsoda and V. Rodríguez-Doncel, Language resources as linked data for the legal domain, *Knowledge of the Law in the Big Data Age* **317** (2019), 170.

[172] W.W.W. Consortium et al., Best practices for publishing linked data, *W3C Working Group Note* (2014).

[173] D. Gromann, Neural language models for the multilingual, transcultural, and multimodal Semantic Web, *Semantic Web* **11**(1) (2020), 29–39. doi:10.3233/SW-190373.

[174] T. Lesnikova, J. David and J. Euzenat, Cross-lingual RDF thesauri interlinking, in: *10th international conference on Language resources and evaluation (LREC)*, No commercial editor., Portoroz, Slovenia, 2016, pp. 2442–2449, lesnikova2016a. https://hal.inria.fr/hal-01382099.

[175] V. Lopez, C. Unger, P. Cimiano and E. Motta, Evaluating question answering over linked data, *Journal of Web Semantics* **21** (2013), 3–13. doi:https://doi.org/10.1016/j.websem.2013.05.006.

[176] J.P. McCrae, P. Cimiano, V.R. Doncel, D. Vila-Suero, J. Gracia, L. Matteis, R. Navigli, A. Abele, G. Vulcu and P. Buitelaar, Reconciling Heterogeneous Descriptions of Language Resources, in: *Proc. of 4th Workshop on Linked Data in Linguistics (LDL'15) at ACL-IJCNLP 2015*, Association for Computational Linguistics (ACL), 2015, pp. 39–42. doi:10.18653/v1/W15-4205.

[177] J.P. McCrae and P. Cimiano, Linghub: a Linked Data based portal supporting the discovery of language resources., *SEMANTiCS (Posters & Demos)* **1481** (2015), 88–91.

[178] K. Moerth, T. Declerck, P. Lendvai and T. Váradi, Accessing Multilingual Data on the Web for the Semantic Annotation of Cultural Heritage Texts., in: *MSW 2011: Multilingual Semantic Web 2011*, Vol. 755, E. Montiel-Ponsoda, J. McCrae, P. Buitelaar and P. Cimiano, eds, CEUR Workshop Proceedings, 2011, pp. 80–85. ISSN 1613-0073.

[179] A.C. Schalley, TYTO–a collaborative research tool for linked linguistic data, in: *Linked Data in Linguistics*, C. Chiarcos, S. Nordhoff and S. Hellmann, eds, Springer, 2012, pp. 139–149. doi:10.1007/978-3-642-28249-2_14.

[180] S. Nordhoff, Linked Data for Linguistic Diversity Research: Glottolog/Langdoc and ASJP Online, Springer Berlin Heidelberg, Berlin, Heidelberg, 2012, pp. 191–200. doi:10.1007/978-3-642-28249-2_18.

[181] M. Ionov, APiCS-Ligt: Towards Semantic Enrichment of Interlinear Glossed Text, 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. 27:1–27:8. ISSN 2190-6807. ISBN 978-3-95977-199-3. doi:10.4230/OASIcs.LDK.2021.27.

[182] F. Gillis-Webber, S. Tittel and C.M. Keet, A model for language annotations on the Web, in: *Knowledge Graphs and Semantic Web: Iberoamerican Knowledge Graphs and Semantic Web Conference*, B. Villazón-Terrazas and Y. Hidalgo-Delgado, eds, Springer, 2019, pp. 1–16. doi:10.1007/978-3-030-21395-4_1.

[183] R. Forkel, The Cross-Linguistic Linked Data Project, in: *Proceedings of the 3nd Workshop on Linked Data in Linguistics (LDL 2014), Reykjavik, Iceland*, 2014, pp. 60–66.

[184] R. Forkel, J.-M. List, S.J. Greenhill, C. Rzymski, S. Bank, M. Cysouw, H. Hammarström, M. Haspelmath, G.A. Kaiping and R.D. Gray, Cross-Linguistic Data Formats, advancing data sharing and re-use in comparative linguistics, *Scientific data* **5**(1) (2018), 1–10.

[185] C. Rzymski, T. Tresoldi, S.J. Greenhill, M.-S. Wu, N.E. Schweikhard, M. Koptjevskaja-Tamm, V. Gast, T.A. Bodt, A. Hantgan, G.A. Kaiping et al., The Database of Cross-Linguistic Colexifications, reproducible analysis of cross-linguistic polysemies, *Scientific data* **7**(1) (2020), 1–12.

[186] L. Sagart, G. Jacques, Y. Lai, R.J. Ryder, V. Thouzeau, S.J. Greenhill and J.-M. List, Dated language phylogenies shed light on the ancestry of Sino-Tibetan, *Proceedings of the National Academy of Sciences* **116**(21) (2019), 10317–10322.

[187] R. Navigli, BabelNet and Friends: A manifesto for multilingual semantic processing, *Intelligenza Artificiale* **7** (2013), 165–181.

[188] F. Bond, C. Fellbaum, S.-K. Hsieh, C.-R. Huang, A. Pease and P. Vossen, A multilingual lexico-semantic database and ontology, in: *Towards the Multilingual Semantic Web*, P. Buitelaar and P. Cimiano, eds, Springer, 2014, pp. 243–258. doi:10.1007/978-3-662-43585-4_15.

[189] M. Ehrmann, G. Jacquet and R. Steinberger, JRC-Names: Multilingual entity name variants and titles as Linked Data, *Semantic Web* (2017). doi:10.3233/SW-160228.

[190] 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*, European Language Resources Association (ELRA), Marseille, France, 2020, pp. 8–15. ISBN 979-10-95546-64-1. https://www.aclweb.org/anthology/2020.iwltp-1.2.

[191] A. Stellato, M. Fiorelli, A. Turbati, T. Lorenzetti, W. van Gemert, D. Dechandon, C. Laaboudi-Spoiden, A. Gerencsér, A. Waniart, E. Costetchi and J. Keizer, VocBench 3: A collaborative Semantic Web editor for ontologies, thesauri and lexicons, *Semantic Web* **11**(5) (2020), 855–881. doi:10.3233/SW-200370.

[192] J. McCrae, E. Montiel-Ponsoda and P. Cimiano, Collaborative semantic editing of linked data lexica, in: *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, N. Calzolari, K. Choukri, T. Declerck, M.U. Doğan, B. Maegaard, J. Mariani, A. Moreno, J. Odijk and S. Piperidis, eds, European Language Resources Association (ELRA), Istanbul, Turkey, 2012, pp. 2619–2625. http://www.lrec-conf.org/proceedings/lrec2012/pdf/544_Paper.pdf.

[193] C. Chiarcos, Get! Mimetypes! Right!, in: *3rd Conference on Language, Data and Knowledge (LDK 2021)*, Vol. 93, D. Gromann, G. Sérasset, T. Declerck, J.P. McCrae, J. Gracia, J. Bosque-Gil, F. Bobillo and B. Heinisch, eds, Schloss Dagstuhl- Leibniz-Zentrum fur Informatik GmbH, Dagstuhl Publishing, 2021, pp. 5:1–5:4. doi:10.4230/OASICS.LDK.2021.5.

[194] L. Heling and M. Acosta, Cost-and robustness-based query optimization for linked data fragments, in: *The Semantic Web – ISWC 2020*, J.Z. Pan, V. Tamma, C. d'Amato, K. Janowicz, B. Fu, A. Polleres, O. Seneviratne and L. Kagal, eds, Springer, 2020, pp. 238–257. doi:10.1007/978-3-030-62419-4_14.

[195] D. Ramos-Vidal and G. de Bernardo, Tool for SPARQL Querying over Compact RDF Representations, *Engineering Proceedings* **7**(1) (2021), 33.

[196] J.P. McCrae, P. Labropoulou, J. Gracia, M. Villegas, V. Rodríguez-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: *Proc. of 12th Extended Semantic Web Conference (ESWC 2015) Satellite Events, Portorož, Slovenia*, Vol. 9341, Springer International Publishing, 2015, pp. 271–282. ISBN 9783319256382. doi:10.1007/978-3-319-25639-9_42.

[197] E.W. De Luca and I. Dahlberg, Including Knowledge Domains from the ICC into the Multilingual Lexical Linked Data Cloud, *Advances in Knowledge Organization* **14** (2014), 258–265.

[198] D. Vila Suero, V. Rodríguez Doncel, A. Gómez-Pérez, P. Cimiano, J.P. McCrae and G. Aguado de Cea, 3LD: Towards high quality, industry-ready linguistic Linked Licensed Data (2014). https://pub.uni-bielefeld.de/download/2732761/2732762/Towards3LD.pdf.

[199] M. Blume, I. Barrière, C. Dye and C. Kang, Challenges for the Development of Linked Open Data for Research in Multilingualism, *Development of Linguistic Linked Open Data Resources for Collaborative Data-Intensive Research in the Language Sciences* (2019), 185–200. doi:10.7551/mitpress/10990.003.0012.

[200] T. Lesnikova, J. David and J. Euzenat, *Algorithms for cross-lingual data interlinking*, hal.archives-ouvertes.fr, 2015. https://hal.archives-ouvertes.fr/hal-01180928/.

[201] J. Gracia, E. Montiel-Ponsoda and A. Gómez-Pérez, Cross-lingual linking on the multilingual web of data (position statement), in: *Proceedings of the 3rd Workshop on the Multilingual Semantic Web (MSW 2012) at ISWC 2012, Boston, USA*, Vol. 936, P. Buitelaar, P. Cimiano, D. Lewis, J. Pustejovsky and F. Sasaki, eds, CEUR Workshop Proceedings, 2012. ISSN 1613-0073.

[202] C. Meilicke, R. García-Castro, F. Freitas, W.R. van Hage, E. Montiel-Ponsoda, R.R. de Azevedo, H. Stuckenschmidt, O. Svab-Zamazal, V. Svatek, A. Tamilin, C. Trojahn and S. Wang, MultiFarm: A Benchmark for Multilingual Ontology Matching, *Web Semantics: Science, Services and Agents on the World Wide Web* **15**(3) (2012). http://www.websemanticsjournal.org/index.php/ps/article/view/315.

[203] J. Gracia, B. Kabashi and I. Kernerman, Results of the Translation Inference Across Dictionaries 2021 Shared Task, in: *Proc. of LDK Workshops and Tutorials 2021*, Vol. 3064, 2021, pp. 208–220. https://tiad2019.unizar.es.

[204] J. Bosque-Gil, V.B. Mititelu, H.G. Oliveira, M. Ionov, J. Gracia, L. Rychkova, G.V. Oleskeviciene, C. Chiarcos, T. Declerck and M. Dojchinovsk, Balancing the digital presence of languages in and for technological development. A Policy Brief on the Inclusion of Data of Under-resourced Languages into the Linked Data Cloud, 2021. https://nexuslinguarum.eu/results/policy-briefshttps://nexuslinguarum.eu/wp-content/uploads/2021/11/PolicyBrief_Under-resourced-languages_SimpleFormat.pdf.

[205] G. de Melo, Lexvo.org: Language-related information for the Linguistic Linked Data cloud, *Semantic Web* **6**(4) (2015), 393–400. doi:10.3233/SW-150171.

[206] SRIA Editorial Team, Strategic Research and Innovation Agenda for the Multilingual Digital Single Market, 2016. http://www.cracking-the-language-barrier.eu/wp-content/uploads/SRIA-V0.9-final-online.pdf.

[207] M.P. di Buono, H.G. Oliveira, V.B. Mititelu, B. Spahiu and G. Nolano, Paving the Way for Enriched Metadata of Linguistic Linked Data, *Sematic Web Journal [under review]* (2022).

[208] A. Gómez-Pérez, D. Vila-Suero, E. Montiel-Ponsoda, J. Gracia and G. Aguado-de-Cea, Guidelines for Multilingual Linked Data, in: *Proceedings of the 3rd International Conference on Web Intelligence, Mining and Semantics*, WIMS '13, Association for Computing Machinery, New York, NY, USA, 2013. ISBN 9781450318501. doi:10.1145/2479787.2479867.

[209] F. Ducel, K. Fort, G. Lejeune and Y. Lepage, Do we Name the Languages we Study? The #BenderRule in LREC and ACL articles, in: *Proceedings of the Language Resources and Evaluation Conference*, European Language Resources Association (ELRA), Marseille, France, 2022, pp. 564–573. https://aclanthology.org/2022.lrec-1.60.