Multilingual Knowledge Graphs and Low-Resource Languages: A Review

Lucie-Aimée Kaffee[1]
Hasso-Plattner Institut, Potsdam, Germany

Russa Biswas
Hasso-Plattner Institut, Potsdam, Germany

C. Maria Keet
University of Cape Town, South Africa

Edlira Kalemi Vakaj
Birmingham City University, UK

Gerard de Melo
Hasso-Plattner Institut, Potsdam, Germany
University of Potsdam, Germany

Abstract

There is a lack of multilingual data to support applications in a large number of languages, especially for low-resource languages. Knowledge graphs (KG) could contribute to closing the gap of language support by providing easily accessible, machine-readable, multilingual linked data, which can be reused across applications. In this paper, we provide an overview of work in the domain of multilingual KGs with a focus on low-resource languages. We review the current state of multilingual KGs along with the different aspects that are crucial for creating KGs with language coverage in mind. Special consideration is given to challenges particular to low-resource languages in KGs. We further provide an overview of applications that yield multilingual KG information as well as downstream applications reusing such multilingual data. Finally, we explore open problems regarding multilingual KGs with a focus on low-resource languages.

2012 ACM Subject Classification Computing methodologies → Natural language processing; Computing methodologies → Semantic networks

Keywords and phrases knowledge graphs, multilingual, low-resource languages, review

Digital Object Identifier 10.4230/TGDK.1.1.10

Category Vision

Received 2023-06-30 **Accepted** 2023-11-17 **Published** 2023-12-19

Editors Aidan Hogan, Ian Horrocks, Andreas Hotho, and Lalana Kagal

Special Issue Trends in Graph Data and Knowledge

1 Introduction

For a wide range of applications, including chatbots and search engines, it is important to support a large variety of languages, as this enables more people to access these applications in their native language. However, currently, many applications only provide support for a highly restricted number of languages. While there are over 7,000 languages spoken in the world[2], applications such as Amazon Alexa or Google Home support 8 or 16, respectively, at the time of writing.[3]

[1] Corresponding author
[2] `https://www.ethnologue.com/`
[3] Link to Amazon Alexa; Link to Google Home.

This drastically limits the accessibility for speakers of languages not among this small set of languages. Providing access to information across languages may be crucial whenever one wishes to make high-quality factual information available to people across the globe. However, for most of the world's languages, the lack of coverage in terms of available data, linguistic models (be they rules-based or data-driven), and the broader tooling ecosystem is daunting, and, therefore, there is insufficient information readily available to easily extend existing applications to the language. We refer to such languages as *low-resource languages*. There have been efforts to make low-resource languages more accessible, as exhibited by initiatives such as Masakhane [27] and SaDiLaR[4], which have created datasets that aim to assist in making language applications available across a greater number of languages. However, the vast majority of language resources are in and for English. For example, on the web at large, 58.8% of websites are estimated to be in English[5], which is also reflected in the size of available corpora[6]. Low-resource languages are barely represented in these sorts of collections. Therefore, models trained on web data are prone to suffering from a severe lack of information in the majority of languages.

One possible way of addressing the issue of a lack of multilingual data is to rely on knowledge graphs (KGs), which store knowledge as graph-structured data [35]. In a data-to-text generation approach, KGs can be used as a source of information for newly generated text across languages [50]. The central storage of language-agnostic information enables downstream applications to provide knowledge, such as in the form of text, for a wide range of language communities. We describe some of the downstream use-cases of multilingual KGs in Section 4.2.

Despite being machine-readable, knowledge graphs also harbour substantial natural language information. Entities and relationships in a knowledge graph generally have natural language labels and often also natural language descriptions. Indeed, among the most prominent knowledge graphs, many provide such natural language labels and descriptions in a multitude of different languages and are thus also a valuable direct source of multilingual data [49]. For instance, a KG may capture that the chemical element gold is called *gold* in English, *altın* in Turkish, *igolide* in isiXhosa, *bulawan* in Cebuano, and so on. It becomes more challenging when there is no simple 1:1 mapping, such as pet being lexicalized as *pet* in English but de facto only described in isiZulu, for instance, as *isilwane sasekhaya* (an animal that is of the home). Similarly, in agglutinating languages, morphemes are strung together to create a new concept that may be translated as a phrase or sentence, such as *umagwazephindelela* for "one who is not satisfied with a single achievement" (a "persistent fighter")[7] [17].

By storing information in the KG in a machine-readable form, the KG can hold information about entities, such as the gold chemical symbol, irrespective of whether this information exists in each of the languages. This helps eliminate language barriers and ensures information is available even in languages where it might otherwise be absent.

In this paper, we explore the current state of multilingual data in knowledge graphs (Section 2), particular challenges regarding low-resource languages (Section 3), approaches to increase language coverage in KGs (Section 4.1), applications using multilingual KGs (Section 4.2), and finally we propose open questions regarding the multilingual support in KGs at large (Section 5).

---

[4] South African Centre for Digital Language Resources with its language resources repository available at `https://repo.sadilar.org/`.

[5] `https://www.statista.com/statistics/262946/most-common-languages-on-the-internet/`, as of January 2023

[6] e.g.: `https://www.sketchengine.eu/corpora-and-languages/corpus-list/`

[7] From: *u-/uma-* noun prefix for noun class 1a or *u-* noun prefix for noun class 1a + *-ma* (v) "stand", *-gwaza* (v) "stab", "slaughter", verb with *-e* ending subjunctive mood, *-phinda* (v) "repeat", "do again", *-phindelela* (v) "do again and again" (note the *-el-* applicative verb extension to *-phinda*)

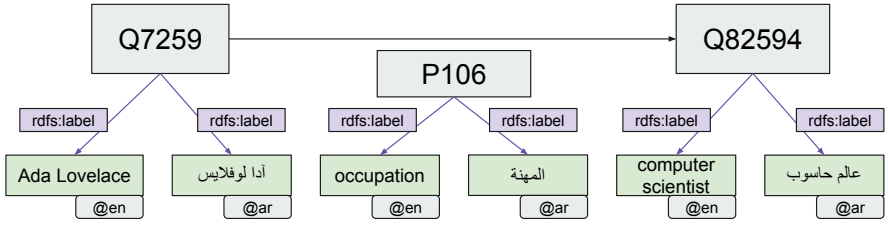

Figure 1 Modelling of multilingual knowledge using a unified knowledge graph with attributes given in multiple different languages, as exemplified by Wikidata. Each entity has a single ID shared across languages. In order to ensure stable and language-neutral IDs, the ID scheme does not include any natural language part intended for human consumption. Instead, a separate entity label is provided for each relevant language (in this example English and Arabic). These are connected using the `rdfs:label` property in the RDF version of Wikidata.

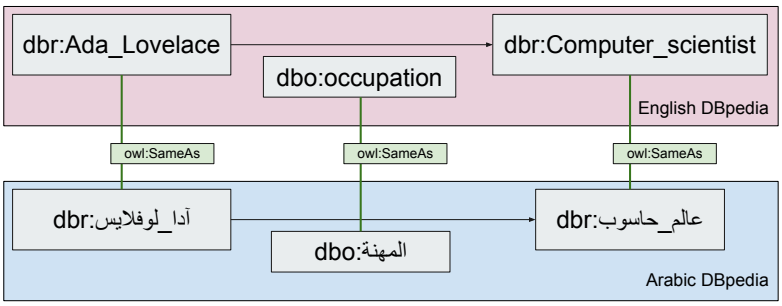

Figure 2 Modelling of multilingual knowledge using separate interconnected entity IDs, as exemplified by DBpedia. A separate entity ID is defined for each language. These are connected by the `owl:sameAs` property, which indicates that they in fact refer to the same entity. In practice, different languages differ widely in their coverage and hence not all information is mirrored across languages as in this example.

2 State of Multilingual KGs

To describe the challenges and opportunities of multilingual KGs, we first provide a brief overview on the current state of knowledge graphs regarding their language coverage. Gracia et al. describe their vision of a multilingual web of data in 2012 as having "*(i) linguistic information for data and vocabularies in different languages, (ii) mappings between data with labels in different languages, and (iii) services to dynamically access and traverse Linked Data across different languages*" [31]. Over 10 years later, we seek to understand what and how much of it has been realised. To gain insights into the state of KGs, we observe multilingual data in KGs from different angles. First, we examine how the different modelling choices and ontologies, i.e., the structure-giving elements of the KG, enable or impede inclusion of multilingual information. Further, it is crucial to assess how many and which languages are supported to what extent. Therefore, we provide an overview of work analysing language coverage across different KGs. Finally, we provide insights on the linguistic information available in KGs in the form of lexicographical data.

2.1 Modelling Multilingual Knowledge

There are, depending on the particular KG, different ways to store multilingual information [29]. A common choice is to assume that KG entities can be shared across languages, and thus for a given KG entity, one can provide relevant language-specific information such as labels in multiple

languages. An example of this is illustrated in Figure 1 for the prominent Wikidata KG. An alternative is to essentially create a different KG (or sub-KG) for each language and then interlink these KGs. An example of this is illustrated in Figure 2 for the DBpedia KG.

The availability of labels can differ substantially across languages. Of course, this may simply be due to incomplete coverage, as we discuss in Section 2.3. In some cases, however, this may stem from linguistic differences in lexicalization or naming: A concept or entity may possess a name in one language but not in another, in which case it is sometimes referred to as a lexical gap. For instance, Runyankore does not have a single word for "pet" but only a description, and isiZulu's *ingcula* does not have a translation into English other than a description of the object (a small bladed hunting spear). At the instance level, additional modelling efforts may be required, if Runyankore-speaking people were to be interested in representing in the KG that Lassie and Scooby Doo are pets, say, or for the development of a KG about Shaka Zulu's armoury.

In the literature, there is a range of guidelines indicating how a KG should represent multilingual information. These include the use of standard ontologies, recommendations for the data itself as well as how the data should be modelled. Across the literature, the following guidelines have been identified, as also summarised in previous work [44].

- **Stable identifiers** [7]: Each entity is identified in the KG with an identifier (ID). These are crucial as the way to access information about an entity. Many authors recommend adopting Unique Resource Identifiers (URIs) or Internationalized Resource Identifiers (IRIs), as used on the Web, for better interoperability. Some KGs identify entities in the graph with identifiers incorporating natural language. For instance, DBpedia uses IRIs such as `http://dbpedia.org/resource/Ada_Lovelace`. Such identifiers are human-readable, and hence easy to interpret for humans, and they also take the function of a label. However, the fact that the natural language portions of such IDs carry meaning can also be a disadvantage. As identifiers are expected to remain stable, such IDs are unable to reflect potential changes in the entity label. For example, if the name of a property changes, the entire structure of the KG would be affected [85]. In contrast, opaque IDs are ones that are not easily readable by humans. For example, they could have a unique identifier for a concept in the form of letters and/or numbers, such as the Wikidata ID `Q7259`. As such IDs do not reflect the label of the entity, if the entity's name changes, the ID is not affected. Moreover, such entity IDs can be more readily shared across languages.
- **Label coverage** [21, 90, 16]: As natural language labels are the way humans access information in the KG and interact with it, it is important that entities, classes, and properties in the KG are labelled. This ensures that information is human-readable and can be displayed to a user.
- **Language tags** [78]: When labelling an entity, it is important to indicate the language in which this label is provided. Even in monolingual KGs, language tags can be valuable, in order to avoid conflating different languages when fusing information from different language sources, especially when automatically merging KGs or when operating across multiple knowledge graphs in different languages through, e.g., federated queries. Language tags can help applications decide which label should be displayed to which user.
- **Language coverage** [31, 44]: For a multilingual KG it is crucial that entities are labelled across a large number of languages for which a relevant label exists. Only such thorough labelling can ensure access of all users to all information, independent of the languages they speak. It also enables KGs to readily make other applications multilingual.
- **Monolingual islands** [31]: When parts of the graph are labelled in only one language, they can form monolingual islands. This can happen because knowledge is prevalent in one cultural context and therefore is yet to be translated, such as governmental initiatives publishing structured data in only their native language. Monolingual islands can lead to worse access to a diverse set of knowledge across the graph and should be addressed when working on a multilingual KG.

- **Reusing existing vocabulary**: In the creation of a KG, it is essential to reuse existing vocabulary, or ontology, to describe the schema of the KG. Especially for label and language information in a KG, it is crucial for easier integration of multiple, complementary KGs. For more information on ontologies, see Section 2.2.
- **Unambiguity** [21]: As there are multiple complementary as well as overlapping ontologies, it is crucial to make sure to avoid ambiguity. According to the Semantic Web standards, it is recommended to use the labelling property `rdfs:label` to provide natural language labels of entities. Furthermore, it is often recommended to provide a single preferred label in a given language per entity, while using other properties such as *alias* to describe alternative names for a concept.

Conceptual differences between different languages may lead to modelling challenges. An important example is the subclass/superclass relationship (which corresponds to hypo-/hypernymy as a lexical relation). For instance, consider the single concept and word for river in English, whereas in the French language and presumably also the corresponding conceptualisation, one distinguishes rivers that flow into other rivers and those that flow into the sea (*fleuve* versus *rivière*) [62]. This may occur similarly for relational properties [54]. A common solution is to treat incompatible concepts across different languages as distinct entries in the KG. Thus, one can avoid conflating the English concept of river with the two French ones and optionally also explicitly describe how the different entries relate to one another cross-lingually. It may also be the case that in one language only a verbal form of a concept exists, i.e., it is assumed to be only a relational property, and in another language it is nominalised, i.e., exists as a unary object or object type only, such that heterogeneous mappings may be needed [26].

2.2 Multilingual Ontologies

Ontologies define the structure of a KG by setting standards for the different properties or relationships to be used, and, for example, the classes used in a KG and across the web of data [36]. The W3C Web Ontology Language (OWL) is a particularly prominent formalism to define ontologies across different KGs to ensure the interoperability of different ontologies [37].

To be able to create multilingual KGs, it is crucial to understand how multilingualism is addressed in the ontologies, be it in the formalism or in a declarative model associated with it, how natural languages are incorporated, and how translation of entities may be recorded.

Gillis-Webber and Keet [29] survey multilinguality in and for ontologies. They grouped the literature and practices in multilingual ontologies into three types of modelling multilingual information: (1) using multilingual labels, i.e., labelling one entity across languages using a labelling property and mentioning the language of the label in the label string (such as in Wikidata in the example in Figure 1), (2) linguistic models (covered in Section 2.4), and (3) mapping-based approach, i.e., creating one entity per language and linking them across languages to each other with an appropriate property (such as in DBpedia in the example in Figure 2). They observed that there is currently a limited uptake of multilingual ontologies at large. Further, limitations of each of the three described ways of modelling multilingual data, such as "accurate representation of languages that require grammatical features such as inflected forms and gender" [30], are yet to be addressed, which apply at least in part also to KGs (discussed in Section 3.2 below).

As ontologies can be defined with the creation of each new KG anew, a number of monolingual ontologies in different languages exist that could be interlinked so as to be more interoperable. Efforts in this direction align ontologies across languages, making it possible to link different KGs across languages [41] or translate labels across languages to make ontologies interoperable and reusable [22].

2.3 Language Coverage Across KGs

Covering a large number of languages in the data in any KG is crucial to be able to support a wide range of languages in downstream applications. For example, in the domain of Question Answering (QA), using a multilingual KG facilitates easy switching between languages and finding the KG best suited for the language requirements of an application and its user base [46].

A range of studies describes the language coverage of KGs to understand how well they currently serve multilingual users. Studies have found that, across numerous language graphs, there is a lack of non-English information in the form of labels. Further, the most used labelling property is `rdfs:label`, and across the web of data at large there is a widespread lack of labels, i.e., a large share of entities are not labelled at all. Additionally, most entities are labelled in only one language [49, 21, 44].

Ell et al. [21] and Kaffee and Simperl [49] developed frameworks to analyse language and label coverage in the web of data, in their case a collection of KGs available online. Zaveri et al. [90] survey data quality metrics and describe a metric for *human readable labelling*, which characterises the coverage of entities by labels. This metric was picked up and made actionable by Debattista et al. [16].

Wikidata specifically shows a slightly different distribution than the web of data at large. While there is still an English bias in the KG, there is a higher degree of language diversity overall [47, 44]. This more varied representation of languages could have different potential causes; for one, there is a multilingual community editing the KG, which leads to a larger number of perspectives added to the graph, compared to KGs that are automatically assembled or mostly contributed to by a single community situated in one part of the world [48]. This international community has been recruited to a large extent from the famous sister project of Wikidata, Wikipedia, where an international community already edits knowledge in their respective language [79]. Further, the community of humans is supported by so-called *bots*, which are automated tools that import and edit knowledge on Wikidata across languages [45]. These approaches can be instructive for future projects seeking to create more multilingual KGs.

This opens promising directions for the future development of multilingual KGs. With a dedicated community and specialised tools, better language coverage can be achieved in the future. In Section 4.1, we discuss automated solutions to improve the language coverage of KGs.

2.4 Lexicographical Data

Given that KGs can be used to store diverse kinds of knowledge, their versatile graph structure also facilitates storing information *about* language. Some KGs, such as Lexvo [14] and Glottolog [33], focus specifically on linguistic metadata about languages and dialects as well as scripts and characters.

Lexicographical data in KGs describes a subset of data that expresses information about the lexicon of languages, i.e., information about words, phrases, and other linguistic expressions. For example, this linguistic information could describe how to conjugate verbs across different languages. In this example, the verb may be modelled as an entity in the graph that is connected to its different grammatical forms through its properties (edges in the graph).

To express and standardise this type of linguistic information, *OntoLex Lemon*[8] was introduced based on the RDF standard, focusing on expressing linguistic information as Linguistic Linked Data [64].

[8] `https://www.w3.org/2016/05/ontolex/`

The concept of lexemes, a form of describing lexical knowledge, has also been introduced to the KG Wikidata. Wikidata uses a custom community-created schema that only loosely resembles Lemon. However, considering the actual data provided by Wikidata, statistics reveal that there are only a few languages with significant lexeme coverage, while the vast majority of languages have little to no representation [73]. This limited language coverage shows that there is still a long way to go to reuse such linguistic information for applications focusing on low-resource languages.

One of the most successful, widely used knowledge resources is WordNet [65], as it has over the years seen extensive use in a number of NLP applications, such as text summarisation [5] and text categorisation [20]. A large number of similar resources have been created for a multitude of different languages, including low-resource languages, and many of these resources have been interlinked. The Universal WordNet [15] and BabelNet [72] were among the first massively multilingual knowledge graphs, both drawing on WordNet as their backbone. WordNet is based on the notion of synonym sets as linguistic concepts, which are connected by various linguistically inspired semantic relationships.

Despite the widespread use of multilingual lexicographic KGs in NLP, it is important to acknowledge that the coverage is uneven. While high-resource languages are well-represented, many widely-spoken languages of socio-economic importance are covered inadequately. For some low-resource languages, only very basic terminology is covered. Many others are missing entirely. While information about grammar can be useful for languages with sparse training data, the lack of representation in KGs cannot yet fill this gap. However, as argued previously, having a central general-domain storage of this information for low-resource languages can support future applications in broadening language coverage. Therefore, we argue, it would be beneficial to build these resources and maintain and widen existing ones.

3 KG Challenges Regarding Low-resource Languages

For KGs to be part of the solution for low-resource languages (LRL) in data-driven settings, they need to be buildable and deployable in KG-driven information systems. They may also need a better specification of "low-resourced languages". With a full characterisation under way [55], within NLP, it is typically narrowed down to LRLs having just limited online corpora, tools, and computational grammars, or lacking something to build statistics-based NLP applications[9]. LRLs have also been characterised as "less studied, resource scarce, less computerised, less privileged, less commonly taught, or low density, among other denomination" [87], and similarly by others [80, 34]. Some studies have sought to quantify this by counting labelled and unlabelled corpora [42] or other data [34] and tools [6, 58], and conducted audits [67, 82, *inter alia*]. Such characterisations are based mostly on quantities, while neglecting to account for the practicalities of working with limited resources. Such practicalities may be grouped into two: one related to the tools and processes and the other at a "deeper" level on language features of both the representation language and the LRL. We'll discuss each in turn.

3.1 Computational Resources

Within the narrow computational resource-oriented view, a consequence is the existence of blocking interdependencies. As a concrete example related to Figure 1, within Abstract Wikipedia [89], it has been proposed to automatically induce templates for a template-based approach [32] or grammars for the Grammatical Framework-based approach [81], ultimately to facilitate rule-based natural language generation to generate Wikipedia articles from the KG. This, however,

[9] See, for instance, Felix Laumann's discussion post at `https://medium.com/neuralspace/low-resource-language-what-does-it-mean-d067ec85dea5`, likely based on Tsvetkov (2017) [87] (slide 26) and repeated in the literature (e.g., [59]).

presupposes the existence and usability of corpora, of good quality part-of-speech taggers and of morphological analysers, which are rarely available for LRLs. These resources are likewise needed for automated KG construction and use in KG-driven information systems, such as educational question generation and document navigation. Consider KG verbalisation for, e.g., isiZulu, a LRL in South Africa spoken by around 23 million people [53]: nouns had to be pluralised, but there was no pluraliser, and verbs of the object properties needed to be conjugated, yet grammar books were outdated and recent linguistics research is scattered, so all that had to be investigated alongside the actual KG task [8, 52]. Also, popular multilingual realisers, such as SimpleNLG [28], are easily adapted among a selection of well-resourced languages in the Indo-European languages family, but they may not be suitable for a language that needs subword-level management, and so a new modular realisation engine may be needed [60]. Conversely, any prospective KG task helps focus language resource development on a specific, measurable, and achievable segment, which is confidence-building and a way to gradually expand the resources.

These anecdotes, and similar observations, are illustrative of several general issues when creating KGs or using them in applications with LRLs, be this for data-oriented techniques and applications or other KG tasks, being:

- There will be linguistic hurdles (gaps in the linguistic knowledge and sociolinguistics) to overcome, which are on top of the intended KG task or KG as a solution. In addition, computerising the language information takes time and scarce human resources, delaying the KG task.
- Freely available pre-existing data is often imprecise, incorrect, or outdated, and thus not a good basis to rely on [solely/at all], requiring an additional data collection stage in a data-driven KG task; therefore, an expert-driven rules-based approach may be more viable for some tasks.
- A resource shortage also tends to imply a human shortage, both in numbers and capacity/-knowledge/skills of the language, limiting the scale of human evaluation and quality of survey or crowdsourced responses.
- The notion of "good" quality is relative and a lower overall size or quality may still mean that the KG task itself performs well (but lower in context due to compounding of less well performing preceding steps).
- Each LRL has its own set of hurdles, and its own history, and how the low resourcedness came about and therewith may need a context-specific incentive to realise the KG task for the LRL.

Also, and separate to KG building, there are power dynamics. Those who build the KG wield power over those who use it and, as Vang argues, "to some degree contests the autonomy of the user" [43]. A multilingual KG would ideally be built *with* the community of prospective users who speak that LRL, not just *for* them. In addition, it is not clear how the notion of KG co-ownership or KG benefit-sharing of extracted and systematised community knowledge, as alluded to in, among others, the groundbreaking San Code of Research Ethics[10], can be realised. Finally, the LRL may have features that do not fit well with the KG language. Since this may be applicable to a subset of LRLs only, we elaborate on this in the next section.

3.2 KG Representation Language Assumptions and Challenges

In this paper, we introduce KGs as a well-suited technology to address limitations in language coverage in downstream applications. There are obvious limitations to this claim. Popular KGs tend to be Eurocentric, in terms of their editors (see for example Wikidata [48]) and data covered.[11]

[10] `https://www.globalcodeofconduct.org/affiliated-codes/`

[11] Visualisation of Wikidata entities with geolocation: `https://wmde.github.io/wikidata-map/dist/index.html`

Moreover, they have been developed to a large extent by researchers working in what is known as the *global north*, emphasising the bias in the type of content and how the content is captured and displayed.

The subject-predicate-object paradigm of capturing facts in a knowledge graph, OWL's assumptions about axiomatisations, and the likes of frame-based approaches suit English well, as it typically takes an object-based approach, has limited verb inflections and has a disjunctive orthography, and an SVO ordering of sentences that are statements.

The practices of modelling assume nouns in the singular and predicates in the 3rd person singular, whose string remains fixed irrespective of the domain or range of the property. This is not the case for several LRLs, however, such as at least Niger-Congo B languages [53], nor for several Indo-European languages that have an extensive noun class system where the verb changes contextually depending on the actor. For instance, the "eats" is all the same in ⟨`Human, eats, apple`⟩, ⟨`Elephant, eats, apple`⟩, and ⟨`Microbe, eats, apple`⟩, regardless who or what (in sg.) is doing the eating, where the "eats" property is being reused as it should be. Consider now a direct translation to isiZulu, respectively: ⟨`Umuntu, udla, i-apula`⟩, ⟨`Indlovu, idla, i-apula`⟩, and ⟨`Igciwane, lidla, i-apula`⟩. The 3rd person singular differs. It is not the case that the predicate is different, just the natural language rendering of it is. However, most KG languages by design typically conflate elements with surface rendering [25]. One could use an identifier, as OBO did, and carry over to OBO Foundry ontologies and in Wikidata, but that still requires additional machinery somewhere to complete the *-dla* verb stem in accordance with the noun class of the noun. Or: to start properly with KG development, one first needs to figure out some sort of extension of, or addition to, RDF. Depending on the LRL, there may well be up to 20 variants, with one for each noun class.

It has also become common practice in "English KGs" to insert prepositions into the property name or label, such as "works *for*" and "part *of*". They may be realised differently in many other languages, such as being affixed to the noun of the class in the range (object) position or infixed in the verb. The affixation to the noun is also specific to the context where it is used, i.e., in which axiom, not the name of the class or individual, for which there is no established KG language yet. An extreme case is the containment relation, typically used in KGs as `contained in`. The notion of containment in isiZulu is realised through indicating the container, by means of noun affixation for locatives and determined by both the container and the containee such that there is no "contained in" verb or name to put in the predicate position [51]; e.g., a bolus of food (*indilinga yokudla*) that is contained in the stomach (*isisu*) becomes *indilinga yokudla isesiswini*. What would the triple be? ⟨`bolus of food, contained in, stomach`⟩ maps neatly with the natural language in English and many other well-resourced languages, but neither a ⟨`indilinga yokudla, blank_prop, isisu`⟩ nor a ⟨`indilinga yokudla, inverse(L3951-S4), isisu`⟩ are satisfactory solutions.

Therefore, a KG may need to be accompanied by a declarative language model and a set of grammar rules, or a different way of usage to represent all the required inputs, or the KG representation language may need to be revised. This is regardless of the usage scenario, from rendering the content of the KG correctly and understandable to the user to automated KG creation. The W3C community standard OntoLex-Lemon [13] as well as the more expressive *lemon* model [61] that aim to provide such a declarative model address this only in part [29, 10].

These language differences lay bare certain English-oriented assumptions baked into the representation language and in naming conventions. This is complicated further if the conceptualisation or terminology diverges not only for domain knowledge but also for a foundational relation such as parthood (e.g., [23, 54]). This need not impede KGs as a possible contributing solution to LRL applications but is to be borne in mind in both KG construction and in their use.

These limitations need to be addressed to make KGs truly inclusive of a wide range of languages.

4 Applications of Multilingual KGs

Multilingual KGs have emerged as resources that transcend language barriers to enable a wide range of applications in our increasingly linked and multilingual world. Improving language coverage in multilingual KGs is critical for completing these resources and making them more relevant to a wide range of global audiences. Also, they have a wide range of applications across various domains due to their ability to bridge linguistic gaps. This section provides an overview of the existing research in the application areas of multilingual KGs.

4.1 Improving Language Coverage of KGs

Recent research has seen an enormous development in the use of KGs to improve Natural Language Processing tasks such as Natural Language Inference (NLI), Question Answering (QA), and Recommender Systems. Even well-known KGs like DBpedia and Wikidata, which are widely used, are the largest in their English versions despite major human efforts to make them available across languages. Furthermore, geographic area-specific facts are frequently restricted to the KG unique to the region or the native language. The incorporation of Machine Learning (ML) models into several languages is constrained by the scarcity of multilingual knowledge.

Machine Translation (MT) systems have been used [2, 63] to improve the language coverage in multilingual KGs, but these efforts only aim to translate domain-specific KGs from English into a target language. These methods ignore the graph structure of KGs, which is critical for determining the domain in which the word must be translated in the target language. Taking into account the graph structure of KGs can help an MT system identify the proper translation for ambiguous labels. As referred to in the survey [70], Rule-based (RBMT), Example-based (EBMT), and Statistical Machine Translation (SMT) based models have been used in the past to translate Semantic Web Technologies (SWT). However, MT for SWT still remains open research due to: (1) lack of clearly specified object attributes, such as cardinality or reflexiveness, (2) concept blending across thesaurus, vocabulary, and ontology, (3) inaccurate definitions of the domain and range, and (4) using ambiguous annotations [70]. However, domain-specific terms from the medical and financial areas have been translated using a Neural Machine Translation (NMT) architecture, outperforming SMT results [3]. Feng et al. [24] introduced a gated NN strategy for translating English KGs into Chinese, learning continuous triple representations. Source and target triples were mapped in the same semantic vector space using their method, which was extracted from Freebase. Their modified NN strategy increased translation accuracy compared to a strong NMT baseline, highlighting the significance of taking into account KG structure for KG translation and enhancing the quality of disambiguation for ambiguous phrases. Another NMT model, THOTH [69] trains to translate the facts from one language into another, treating the facts (i.e., triples) in a KG as sentences with URIs acting as tokens. It uses KG embeddings and two separate recurrent neural network models to extract bilingual alignments between a source and target KG and then learns the translation.

KG embedding based models have been proposed to enhance multilingual KGs. A more realistic method would draw on the information in several language-specific KGs, keeping in mind that individual KGs have their own strengths and limitations on data quality and coverage. This is a significant challenge since inconsistently expressed facts and a lack of sufficient alignment information can make it difficult for knowledge to be transferred between many independently maintained KGs. KEnS [11] is one such approach that embeds all multilingual KGs in a shared embedding space where the association of entities is captured through self-learning.

LLM-based KG creation has been used recently in an attempt to create monolingual knowledge facts for KGs [77], but multilingual KG construction and enrichment have not yet been attempted at their full strength. Due to the potential for complementary and unequally dispersed

knowledge stated in many languages, multilingual LLMs can offer richer combined knowledge for multilingual KGs. Prix-LM [91] is a pioneering model for multilingual KG construction and completion that employs both monolingual triples and cross-lingual linkages retrieved from existing multilingual KGs, followed by fine-tuning a multilingual language encoder XLM-R via a causal language modelling objective. Hou et al. [38] proposes another model that leverages adapters to fine-tune LLMs for multilingual KGs, including low-resource languages, enhancing the corresponding downstream tasks. Therefore, leveraging LLMs for the construction and completion of multilingual KGs emerges to be one of the promising future directions of research, considering that their information content has shown promise for knowledge extraction [77].

4.2 Downstream Applications of Multilingual KG Information

In this section, we survey a set of use cases for multilingual KGs. When developing new KGs, new ontologies, or improving the language coverage of existing KGs, it is crucial to keep the use-cases in mind to ensure that the KG can indeed serve knowledge that is useful in the envisioned tasks. Several applications already make use of multilingual KGs, in the following we present a selected few tasks.

Multilingual Knowledge Graph Question Answering (mKGQA) is a task that involves answering a user's questions in a set of languages based on facts stored in a KG. This is a research topic of particular importance, as these systems can bridge the gap between users and pertinent information on the Web. The task should ideally address all languages, making it possible for users to ask questions independent of the languages of the information in the KG. Currently, however, these systems typically define a set of languages in which questions can be answered, relying on information in the corresponding language stored in the target KG. mKGQA is strongly dependent on data, and there is currently a lack of multilingual data on the web, making it accessible to only a fraction of people[12], therefore causing a "cultural gap" on the Web [66]. Further, the task of mKGQA suffers from a lack of multilingual benchmarks.

Perevalov et al. [75, 76] identified 17 mKGQA systems. We have provided in the table below a list of 4 mKGQA systems filtered based on the coverage that they provide for languages besides English (with a focus on lower-resourced languages), along with the KG used, data sets, and the specific languages covered.

Table 1 Multilingual Knowledge Graph Question Answering (mKGQA) systems.

mKGQA system	Ref.	KG	Languages	Dataset
WDAqua-core, 2018	[19]	DBpedia, Wikidata, DBLP, MusicBrainz	en, de, es, it	QALD, LC-QuAD 1.0
QAnswer, 2019	[18]	DBpedia, Wikidata, DBLP, MusicBrainz, FreeBase	en, de, fr, it, es, pt, ar, zh	QALD-3-7, LC-QuAD 1.0
Y. Zhou et al.	[92]	DBpedia	en, fa, de, ro, it, ru, fr, nl, es, hi, pt	LC-QuAD 1.0, QALD-9
A. Perevalov et al.	[74]	Wikidata, DBpedia	en, de, fr, ru, uk, lt, be, ba, hy	QALD-9-Plus
BLEU-4	[84]	–	tibetian	TibetanQA+

[12] 25.9% according to https://www.statista.com/statistics/262946/share-of-the-most-common-languages-on-the-internet/, retrieved 2023-02-15

Knowledge Graph Completion aims to add new, missing facts to a KG. This remains a challenging problem that is particularly pronounced in multilingual low-resource languages, given that human annotations are rare and difficult to procure [39]. A number of solutions have been devised to mitigate this problem, such as using KG embeddings [40] or jointly predicting entity alignment across languages and new facts for KG completion [9, 86, 11]. A promising approach is the self-supervised adaptive graph alignment (SS-AGA) method [39], which regards alignment as a new edge type between parallel entities instead of a loss constraint, and fuses KGs from different languages in a single graph. This approach has been evaluated on both the public multilingual DBpedia KG and a newly created industrial multilingual E-commerce KG.

Cross Lingual Fact Extraction (CLFE) is the task of extracting facts from a text to, e.g., store the facts in a structured data format. Extracting facts from source text of different languages has not received as much attention as monolingual fact extraction [83]. KGs can support multilingual fact linking and extraction, which is important in many downstream tasks such as QA. The REFCOGLink model [57] is based on linking facts expressed in a sentence to the corresponding fact labels (i.e., language-specific representation of the fact) in the KG and outperforms standard retrieval + re-ranking. The CLFE (Cross-Lingual Fact Extraction) [1] model demonstrates strong performance in multilingual and cross-lingual fact extraction tasks, specifically in English and seven other LR Indic languages. It achieves an F1 score of 77.46% using two different approaches. The first approach is a classification-based method, where the model first extracts the object or "tail" of the fact and then predicts the relationship between the extracted tail and the subject. The second approach is a generative one, which combines both tasks into a single step. CLFE makes use of the XAlign dataset, which contains 0.45M pairs across 8 languages, of which 5,402 pairs have been manually annotated.

Multilingual Relation Classification is the task of extracting relations (i.e., triples in the context of KGs) from natural language text in various languages. Multilingual relation classification has been explored through the method of prompting, which can receive promising results even for lower-resourced languages [12]. One of the challenges for the task of multilingual relation classification is the lack of multilingual datasets and benchmarks. The dataset RELX [56] aims to close this gap by providing a baseline model and benchmark for English, French, German, Spanish, and Turkish. IndoRE [71] is a comprehensive dataset comprising 21,000 gold-tagged sentences for named entity recognition (NER) and relation extraction (RE) in three Indian languages (English, Bengali, Hindi) as well as English. The dataset provides valuable resources for advancing research in relation classification of Indian languages, which is crucial for KG augmentation and Question Answering systems. The authors employ multilingual BERT and transfer learning techniques, and propose TransRel, a multilingual system for joint named entity recognition (NER) and relation extraction (RE) with interlingual transfer.

Neural Machine Translation translates text from a source to a target language using deep learning architectures. The performance of Neural Machine Translation (NMT) systems can be enhanced through the integration of KGs, particularly when translating domain-specific expressions and named entities. For example, Moussallem et al. [68] introduce an approach incorporating the KG DBpedia into NMT models, resulting in considerable improvements in performance of these NMT models.

Automatic KG creation for low-resource languages has been leveraged as another possible route to tackle the lack of low-resource language KGs. HKC has been recently introduced as a framework that constructs a knowledge graph for the Hindi language [88] using various NLP techniques. FarsBase, a Persian multi-source knowledge graph [4] is another example of such an application. To construct such a KG, the authors apply a number of techniques to integrate data from Wikipedia and both structured and unstructured data from the web.

Overall, KGs offer immense potential in supporting multilingualism, which in the future will likely also increasingly benefit low-resource languages. By constructing multilingual KGs, aligning entities across languages, and employing techniques such as machine translation and cross-lingual extraction, KGs can effectively bridge the knowledge gap and facilitate access to information in low-resource languages. These advancements not only empower speakers of all languages but also contribute to a more inclusive and diverse knowledge ecosystem.

5 Open Problems

We have introduced existing approaches to close the language gap in the domain of KGs. We described multilingual KGs, their challenges with regard to low-resource languages, and approaches to improve and use them. However, there remain a number of open challenges, which we will describe here to build a foundation for future work and point out some of the pressing issues in the domain of multilingual KGs.

5.1 Regarding the State of Multilingual KGs

The lack of multilingual information in KGs should be addressed in the future, and while there are a number of approaches seeking to mitigate this problem, none have thus far succeeded in addressing this important challenge. One of the notable issues with current KGs is that they are typically not aligned across languages, given the usage of different language ontologies. This misalignment poses a large challenge when merging knowledge across languages. Addressing the cross-lingual alignment of ontologies is one of the crucial challenges that will build the foundation for the wide availability of cross-lingual information.

Merging KGs across languages, even with aligned ontologies, remains a non-trivial task. Not only is there a need to identify alignments at the entity and class level. Future studies are needed to understand which languages are covered to which extent in the different KGs and how these different KGs with potentially different topic-focus could be merged harmonically without centering one language in the approach.

Another open problem, especially for community-edited KGs, is the question of how to interest a larger, more diverse community in the contribution to the knowledge stored in the KG. If a diverse community is contributing, naturally more languages will be covered. Creating incentives and showing how the information could be used, such as building tools based on KG information, could be a way to address this challenge. However, future work will have to better understand existing incentives and broaden them.

5.2 Regarding Low-resource Languages

While there have been approaches suggested to address the lack of languages covered in KGs (see Section 4), there is currently a glaring gap when it comes to low-resource languages. Having very little available information in any language is a challenge for all applications, especially neural- or deep-learning-driven ones, such as large language models. This then also raises the question of optimal strategies for KG construction, such as human-in-the-loop procedures due to limited and overly noisy data and appropriate incentive strategies for manual modelling. KGs could contribute to closing this gap to some extent by providing central storage of multilingual, linked, reusable data. One of the advantages of using KGs may be that while a fact is not yet translated to a lower-resourced language, it already exists in another language, and by linking the data across languages, these facts can be reused even before translation. However, the challenges described in Section 3.2 in the form of assumptions of language modelling need to be urgently addressed

for under-represented languages to be able to catch up in the content representation. From an ontology engineering perspective, it is important to focus on representing the diversity of language structures accurately, lest the technique limits its use to applications in a few languages only rather than the breadth of opportunities across the world.

5.3 Regarding Applications of Multilingual KGs

For the reuse of multilingual KGs in downstream tasks, there are still many opportunities to better integrate multilingual KG information with approaches from the field of NLP and with multi-modal approaches. Reusing the knowledge graph information is not only beneficial to the applications that use them but can also create an incentive to create better, more diverse KGs. To ensure this, the current information needs to be accurately curated, and multilingual information should be put into focus for future approaches in increasing KG coverage. Understanding coverage in different low-resource languages as well as reusing this information is a promising path forward for future work that can have a real-world impact.

6 Conclusion

In this paper, we reviewed different aspects of multilingual KGs; we established the state of multilingual KGs by summarising the guidelines on creating multilingual KGs; established the challenges with regards to low-resource languages; described applications for and using multilingual KGs; and finally pointed out open problems derived from our survey of multilingual KGs. The literature provides clear guidelines and ontologies for multilingual information and lexicographic data. Yet, there is a severe lack of multilingual information in existing KGs and a bias toward English-language information. Particularly for low-resource languages, KGs could be useful for closing information gaps. However, we concluded that there are currently major challenges that need to be addressed regarding low-resource language integration into KGs, such as the English-centric structure and content of existing KGs.

Current approaches to improve language coverage of KGs, such as machine translation of KG labels or leveraging KG embeddings to align monolingual KGs across languages, are a promising direction to make the KG more diverse. However, these approaches are currently limited to a small set of languages and need to be explored for low-resource languages with a focus on the challenges described. Future work has to focus on the inclusion of non-European languages from the very structure of KGs, including them in the considerations of how knowledge is modelled in KGs. Expanding language inclusivity holds immense importance as it paves the way for a more accessible and all-encompassing digital landscape, enabling the internet to cater to a diverse array of communities. While currently language coverage, especially for low-resource languages, is limited, there are viable avenues for progress. If these pathways are pursued, KGs could serve as a technology to realise the vision of a more equitable and inclusive internet, facilitating the exchange of knowledge across language communities.

References

1 Tushar Abhishek, Shivprasad Sagare, Bhavyajeet Singh, Anubhav Sharma, Manish Gupta, and Vasudeva Varma. Xalign: Cross-lingual fact-to-text alignment and generation for low-resource languages. In *Companion Proceedings of the Web Conference 2022*, pages 171–175, 2022. doi:10.1145/3487553.3524265.

2 Mihael Arcan and Paul Buitelaar. Ontology Label Translation. *Human Language Technologies: Conference of the North American Chapter of the Association of Computational Linguistics, Proceedings, June 9-14, 2013, Westin Peachtree Plaza Hotel, Atlanta, Georgia, USA*, pages 40–46, 2013. URL: https://aclanthology.org/N13-2006/.

3 Mihael Arcan and Paul Buitelaar. Translating Domain-Specific Expressions in Knowledge Bases with Neural Machine Translation. *CoRR*, abs/1709.02184, 2017. doi:10.48550/arXiv.1709.02184.

4 Majid Asgari-Bidhendi, Ali Hadian, and Behrouz Minaei-Bidgoli. Farsbase: The persian knowledge graph. *Semantic Web Journal*, 10(6):1169–1196, 2019. doi:10.3233/SW-190369.

5 Kedar Bellare, Anish Das Sarma, Atish Das Sarma, Navneet Loiwal, Vaibhav Mehta, Ganesh Ramakrishnan, and Pushpak Bhattacharyya. Generic Text Summarization Using WordNet. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation, LREC 2004, May 26-28, 2004, Lisbon, Portugal*. European Language Resources Association, 2004. URL: http://www.lrec-conf.org/proceedings/lrec2004/summaries/342.htm.

6 V. Berment. *Méthodes pour informatiser des langues et des groupes de langues peu dotées*. Phd thesis, J. Fourier University – Grenoble I, may 2004. URL: https://theses.hal.science/tel-00006313/document.

7 Tim Berners-Lee. Cool URIs don't change, 1998. Accessed on 05.07.2023. URL: https://www.w3.org/Provider/Style/URI.html.

8 Joan Byamugisha, C. Maria Keet, and Langa Khumalo. Pluralising nouns in isiZulu and related languages. In A. Gelbukh, editor, *Proceedings of CICLing'16*, volume 9623 of *LNCS*, pages 271–283. Springer, 2018. doi:10.1007/978-3-319-75477-2_18.

9 Soumen Chakrabarti, Harkanwar Singh, Shubham Lohiya, Prachi Jain, and Mausam. Joint completion and alignment of multilingual knowledge graphs. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang, editors, *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 11922–11938. Association for Computational Linguistics, 2022. doi:10.18653/V1/2022.EMNLP-MAIN.817.

10 Catherine Chavula and C. Maria Keet. Is lemon sufficient for building multilingual ontologies for Bantu languages? In C. Maria Keet and Valentina Tamma, editors, *Proceedings of the 11th OWL: Experiences and Directions Workshop (OWLED'14)*, volume 1265 of *CEUR-WS*, pages 61–72, 2014. Riva del Garda, Italy, Oct 17-18, 2014. URL: https://ceur-ws.org/Vol-1265/owled2014_submission_10.pdf.

11 Xuelu Chen, Muhao Chen, Changjun Fan, Ankith Uppunda, Yizhou Sun, and Carlo Zaniolo. Multilingual knowledge graph completion via ensemble knowledge transfer. In Trevor Cohn, Yulan He, and Yang Liu, editors, *Findings of the Association for Computational Linguistics: EMNLP 2020, Online Event, 16-20 November 2020*, volume EMNLP 2020 of *Findings of ACL*, pages 3227–3238. Association for Computational Linguistics, 2020. doi:10.18653/V1/2020.FINDINGS-EMNLP.290.

12 Yuxuan Chen, David Harbecke, and Leonhard Hennig. Multilingual relation classification via efficient and effective prompting. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang, editors, *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 1059–1075. Association for Computational Linguistics, 2022. doi:10.18653/V1/2022.EMNLP-MAIN.69.

13 Philipp Cimiano, John P. McCrae, and Paul Buitelaar. Lexicon model for ontologies: Community report. Final community group report, 10 may 2016, W3C, 2016. URL: https://www.w3.org/2016/05/ontolex/.

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

15 Gerard de Melo and Gerhard Weikum. Towards a universal wordnet by learning from combined evidence. In David Wai-Lok Cheung, Il-Yeol Song, Wesley W. Chu, Xiaohua Hu, and Jimmy J. Lin, editors, *Proceedings of the 18th ACM Conference on Information and Knowledge Management, CIKM 2009, Hong Kong, China, November 2-6, 2009*, pages 513–522, New York, NY, USA, 2009. ACM. doi:10.1145/1645953.1646020.

16 Jeremy Debattista, Sören Auer, and Christoph Lange. Luzzu - A methodology and framework for linked data quality assessment. *ACM J. Data Inf. Qual.*, 8(1):4:1–4:32, 2016. doi:10.1145/2992786.

17 G. R. Dent and C. L. S. Nyembezi. *Scholar's Zulu Dictionary*. Shuter & Shooter Publishers, 4 edition, 2009.

18 Dennis Diefenbach, Pedro Henrique Migliatti, Omar Qawasmeh, Vincent Lully, Kamal Singh, and Pierre Maret. QAnswer: A Question Answering prototype bridging the gap between a considerable part of the LOD cloud and end-users. In *The World Wide Web Conference*, pages 3507–3510, 2019. doi:10.1145/3308558.3314124.

19 Dennis Diefenbach, Kamal Singh, and Pierre Maret. WDAqua-core1: A Question Answering service for RDF Knowledge Bases. In *Companion Proceedings of the The Web Conference 2018*, pages 1087–1091, 2018. doi:10.1145/3184558.3191541.

20 Zakaria Elberrichi, Abdellatif Rahmoun, and Mohamed Amine Bentaallah. Using wordnet for text categorization. *Int. Arab J. Inf. Technol.*, 5(1):16–24, 2008.

21 Basil Ell, Denny Vrandecic, and Elena Simperl. Labels in the web of data. In Lora Aroyo, Chris Welty, Harith Alani, Jamie Taylor, Abraham Bernstein, Lalana Kagal, Natasha Fridman Noy, and Eva Blomqvist, editors, *The Semantic Web - ISWC 2011 - 10th International Semantic Web Conference, Bonn, Germany, October 23-27, 2011, Proceedings, Part I*, volume 7031 of *Lecture Notes in Computer Science*, pages 162–176. Springer, 2011. doi:10.1007/978-3-642-25073-6_11.

22 Mauricio Espinoza, Asunción Gómez-Pérez, and Eduardo Mena. Enriching an ontology with multilingual information. In Sean Bechhofer, Manfred Hauswirth, Jörg Hoffmann, and Manolis Koubarakis, editors, *The Semantic Web: Research and Applications, 5th European Semantic Web Conference, ESWC 2008, Tenerife, Canary Islands, Spain, June 1-5, 2008, Proceedings*, volume 5021 of *Lecture Notes in Computer*

Science, pages 333–347. Springer, 2008. doi:10.1007/978-3-540-68234-9_26.

23 Chen-Chieh Feng and David M. Mark. Cross-linguistic research on landscape categories using geonet names server data: A case study for indonesia and malaysia. *The Professional Geographer*, 69(4):567–578, 2017. doi:10.1080/00330124.2017.1288575.

24 Xiaocheng Feng, Duyu Tang, Bing Qin, and Ting Liu. English-Chinese Knowledge Base Translation with Neural Network. *COLING 2016, 26th International Conference on Computational Linguistics, Proceedings of the Conference: Technical Papers, December 11-16, 2016, Osaka, Japan*, pages 2935–2944, 2016. URL: https://aclanthology.org/C16-1276/.

25 P. R. Fillottrani and C. M. Keet. An analysis of commitments in ontology language design. In B. Brodaric and F. Neuhaus, editors, *11th International Conference on Formal Ontology in Information Systems 2020 (FOIS'20)*, volume 330 of *FAIA*, pages 46–60. IOS Press, 2020. doi:10.3233/FAIA200659.

26 Pablo R. Fillottrani and C. Maria Keet. Patterns for Heterogeneous TBox Mappings to Bridge Different Modelling Decisions. In E. Blomqvist et al., editors, *Proceeding of the 14th Extended Semantic Web Conference (ESWC'17)*, volume 10249 of *LNCS*, pages 371–386. Springer, 2017. 30 May - 1 June 2017, Portoroz, Slovenia. doi:10.1007/978-3-319-58068-5_23.

27 ∀, Wilhelmina Nekoto, Vukosi Marivate, Tshinondiwa Matsila, Timi Fasubaa, Tajudeen Kolawole, Taiwo Fagbohungbe, Solomon Oluwole Akinola, Shamsuddee Hassan Muhammad, Salomon Kabongo, Salomey Osei, et al. Participatory Research for Low-resourced Machine Translation: A Case Study in African Languages. *Findings of EMNLP*, 2020. doi:10.18653/v1/2020.findings-emnlp.195.

28 A. Gatt and E. Reiter. Simplenlg: A realisation engine for practical applications. In E. Krahmer and M. Theune, editors, *Proceedings of the 12th European Workshop on Natural Language Generation (ENLG'09)*, pages 90–93. ACL, 2009. March 30-31, 2009, Athens, Greece. URL: https://aclanthology.org/W09-0613.pdf.

29 Frances Gillis-Webber and C. Maria Keet. A review of multilingualism in and for ontologies. *CoRR*, abs/2210.02807, 2022. doi:10.48550/ARXIV.2210.02807.

30 Frances Gillis-Webber and C. Maria Keet. A survey of multilingual OWL ontologies in bioportal. In Katy Wolstencroft, Andrea Splendiani, M. Scott Marshall, Chris Baker, Andra Waagmeester, Marco Roos, Rutger A. Vos, Rianne Fijten, and Leyla Jael Castro, editors, *13th International Conference on Semantic Web Applications and Tools for Health Care and Life Sciences, SWAT4HCLS 2022, Virtual Event, Leiden, The Netherlands, January 10th to 14th, 2022*, volume 3127 of *CEUR Workshop Proceedings*, pages 87–96. CEUR-WS.org, 2022. URL: https://ceur-ws.org/Vol-3127/paper-11.pdf.

31 Jorge Gracia, Elena Montiel-Ponsoda, Philipp Cimiano, Asunción Gómez-Pérez, Paul Buitelaar, and John P. McCrae. Challenges for the multilingual web of data. *J. Web Semant.*, 11:63–71, 2012. doi:10.1016/J.WEBSEM.2011.09.001.

32 Ariel Gutman and C. Maria Keet. Template language for wikifunctions, 2022. URL: https://meta.wikimedia.org/wiki/Abstract_Wikipedia/Template_Language_for_Wikifunctions.

33 Harald Hammarström, Robert Forkel, Martin Haspelmath, and Sebastian Bank. *Glottolog. Version 4.8*. Max Planck Institute for Evolutionary Anthropology, Leipzig, 2023. doi:10.5281/zenodo.8131084.

34 Michael A. Hedderich, Lukas Lange, Heike Adel, Jannik Strötgen, and Dietrich Klakow. A survey on recent approaches for natural language processing in low-resource scenarios. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2545–2568. Association for Computational Linguistics, jun 2021. doi:10.18653/V1/2021.NAACL-MAIN.201.

35 Aidan Hogan, Eva Blomqvist, Michael Cochez, Claudia d'Amato, Gerard de Melo, Claudio Gutierrez, Sabrina Kirrane, José Emilio Labra Gayo, Roberto Navigli, Sebastian Neumaier, Axel-Cyrille Ngonga Ngomo, Axel Polleres, Sabbir M. Rashid, Anisa Rula, Lukas Schmelzeisen, Juan Sequeda, Steffen Staab, and Antoine Zimmermann. *Knowledge Graphs*. Number 22 in Synthesis Lectures on Data, Semantics, and Knowledge. Morgan & Claypool, 2021. doi:10.2200/S01125ED1V01Y202109DSK022.

36 Ian Horrocks. Ontologies and the semantic web. *Communications of the ACM*, 51(12):58–67, 2008. doi:10.1007/3-540-45810-7.

37 Ian Horrocks, Peter F. Patel-Schneider, and Frank van Harmelen. From SHIQ and RDF to OWL: the making of a web ontology language. *J. Web Semant.*, 1(1):7–26, 2003. doi:10.1016/J.WEBSEM.2003.07.001.

38 Yifan Hou, Wenxiang Jiao, Meizhen Liu, Carl Allen, Zhaopeng Tu, and Mrinmaya Sachan. Adapters for Enhanced Modeling of Multilingual Knowledge and Text. *Findings of the Association for Computational Linguistics: EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 3902–3917, 2022. doi:10.18653/V1/2022.FINDINGS-EMNLP.287.

39 Zijie Huang, Zheng Li, Haoming Jiang, Tianyu Cao, Hanqing Lu, Bing Yin, Karthik Subbian, Yizhou Sun, and Wei Wang. Multilingual knowledge graph completion with self-supervised adaptive graph alignment. *arXiv preprint arXiv:2203.14987*, 2022. doi:10.48550/arXiv.2203.14987.

40 Zijie Huang, Zheng Li, Haoming Jiang, Tianyu Cao, Hanqing Lu, Bing Yin, Karthik Subbian, Yizhou Sun, and Wei Wang. Multilingual knowledge graph completion with self-supervised adaptive graph alignment. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio, editors, *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 474–485. Association for Com-

putational Linguistics, 2022. doi:10.18653/V1/2022.ACL-LONG.36.

41 Shimaa Ibrahim, Said Fathalla, Jens Lehmann, and Hajira Jabeen. Toward the multilingual semantic web: Multilingual ontology matching and assessment. *IEEE Access*, 11:8581–8599, 2023. doi:10.1109/ACCESS.2023.3238871.

42 Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. The state and fate of linguistic diversity and inclusion in the NLP world. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293, Online, jul 2020. Association for Computational Linguistics. doi:10.18653/V1/2020.ACL-MAIN.560.

43 K. Juel Vang. Ethics of Google's Knowledge Graph: some considerations. *Journal of Information, Communication and Ethics in Society*, 11(4):245–260, 2013. doi:10.1108/JICES-08-2013-0028.

44 Lucie-Aimée Kaffee. *Multilinguality in Knowledge Graphs*. PhD thesis, University of Southampton, 2021. URL: https://eprints.soton.ac.uk/456783/.

45 Lucie-Aimée Kaffee, Kemele M. Endris, and Elena Simperl. When humans and machines collaborate: cross-lingual label editing in wikidata. In Björn Lundell, Jonas Gamalielsson, Lorraine Morgan, and Gregorio Robles, editors, *Proceedings of the 15th International Symposium on Open Collaboration, OpenSym 2019, Skövde, Sweden, August 20-22, 2019*, pages 16:1–16:9. ACM, 2019. doi:10.1145/3306446.3340826.

46 Lucie-Aimée Kaffee, Kemele M. Endris, Elena Simperl, and Maria-Esther Vidal. Ranking knowledge graphs by capturing knowledge about languages and labels. In Mayank Kejriwal, Pedro A. Szekely, and Raphaël Troncy, editors, *Proceedings of the 10th International Conference on Knowledge Capture, K-CAP 2019, Marina Del Rey, CA, USA, November 19-21, 2019*, pages 21–28. ACM, 2019. doi:10.1145/3360901.3364443.

47 Lucie-Aimée Kaffee, Alessandro Piscopo, Pavlos Vougiouklis, Elena Simperl, Leslie Carr, and Lydia Pintscher. A glimpse into babel: An analysis of multilinguality in wikidata. In Lorraine Morgan, editor, *Proceedings of the 13th International Symposium on Open Collaboration, OpenSym 2017, Galway, Ireland, August 23-25, 2017*, pages 14:1–14:5. ACM, 2017. doi:10.1145/3125433.3125465.

48 Lucie-Aimée Kaffee and Elena Simperl. Analysis of editors' languages in wikidata. In *Proceedings of the 14th International Symposium on Open Collaboration, OpenSym 2018, Paris, France, August 22-24, 2018*, pages 21:1–21:5. ACM, 2018. doi:10.1145/3233391.3233965.

49 Lucie-Aimée Kaffee and Elena Simperl. The human face of the web of data: A cross-sectional study of labels. In Anna Fensel, Victor de Boer, Tassilo Pellegrini, Elmar Kiesling, Bernhard Haslhofer, Laura Hollink, and Alexander Schindler, editors, *Proceedings of the 14th International Conference on Semantic Systems, SEMANTiCS 2018, Vienna, Austria, September 10-13, 2018*, volume 137 of *Procedia Computer Science*, pages 66–77. Elsevier, 2018. doi:10.1016/J.PROCS.2018.09.007.

50 Lucie-Aimée Kaffee, Pavlos Vougiouklis, and Elena Simperl. Using natural language generation to bootstrap missing wikipedia articles: A human-centric perspective. *Semantic Web*, 13(2):163–194, 2022. doi:10.3233/SW-210431.

51 C. M. Keet. Representing and aligning similar relations: parts and wholes in isizulu vs english. In J. Gracia, F. Bond, J. McCrae, P. Buitelaar, C. Chiarcos, and S. Hellmann, editors, *Language, Data, and Knowledge 2017 (LDK'17)*, volume 10318 of *LNAI*, pages 58–73. Springer, 2017. 19-20 June, 2017, Galway, Ireland. doi:10.1007/978-3-319-59888-8_5.

52 C. M. Keet and L. Khumalo. Grammar rules for the isiZulu complex verb. *Southern African Journal of Language and Linguistics*, 35(2):183–200, 2017. doi:10.2989/16073614.2017.1358097.

53 C. Maria Keet and Langa Khumalo. Toward a knowledge-to-text controlled natural language of isiZulu. *Language Resources and Evaluation*, 51(1):131–157, 2017. doi:10.1007/S10579-016-9340-0.

54 C. Maria Keet and Langa Khumalo. Parthood and part–whole relations in zulu language and culture. *Applied Ontology*, 15(3):361–384, 2020. doi:10.3233/AO-200230.

55 C. Maria Keet and Langa Khumalo. Contextualising levels of language resourcedness affecting digital processing of text. *CoRR*, abs/2309.17035, 2023. doi:10.48550/ARXIV.2309.17035.

56 Abdullatif Köksal and Arzucan Özgür. The RELX dataset and matching the multilingual blanks for cross-lingual relation classification. In Trevor Cohn, Yulan He, and Yang Liu, editors, *Findings of the Association for Computational Linguistics: EMNLP 2020, Online Event, 16-20 November 2020*, volume EMNLP 2020 of *Findings of ACL*, pages 340–350. Association for Computational Linguistics, 2020. doi:10.18653/V1/2020.FINDINGS-EMNLP.32.

57 Keshav Kolluru, Martin Rezk, Pat Verga, William W. Cohen, and Partha Talukdar. Multilingual Fact Linking, 2021. doi:10.48550/arXiv.2109.14364.

58 Steven Krauwer. The basic language resource kit (BLARK) as the first milestone for the language resources roadmap. In *Proceedings of the 2003 International Workshop Speech and Computer SPECOM'03*, volume 2003, pages 8–15, 2003. Moscow, Russia, 2003. URL: http://www.elsnet.org/dox/krauwer-specom2003.pdf.

59 A. Magueresse, V. Carles, and E. Heetderks. Low-resource languages: A review of past work and future challenges. *CoRR*, abs/2006.07264, 2020. doi:10.48550/arXiv.2006.07264.

60 Zola Mahlaza. *Foundations for reusable and maintainable surface realisers for isiXhosa and isiZulu*. Phd thesis, Department of Computer Science, University of Cape Town, South Africa, nov 2022. URL: https://adeebnqo.github.io/files/Thesis.pdf.

61 John McCrae, Guadalupe Aguado-de Cea, Paul Buitelaar, Philipp Cimiano, Thierry Declerck, Asunción Gómez-Pérez, Jorge Gracia, Laura Hollink, Elena Montiel-Ponsoda, Dennis Spohr,

and Tobias Wunner. Interchanging lexical resources on the semantic web. *Language Resources and Evaluation*, 46(4):701–719, 2012. doi:10.1007/S10579-012-9182-3.

62 John McCrae, Guadalupe Aguado de Cea, Paul Buitelaar, Philipp Cimiano, Thierry Declerck, Asunción Gómez-Pérez, Jorge Gracia, Laura Hollink, Elena Montiel-Ponsoda, Dennis Spohr, and Tobias Wunner. The lemon cookbook. Technical report, Monnet Project, jun 2012. URL: https://www.lemon-model.net/learn/cookbook.php.

63 John P. McCrae, Mihael Arcan, Kartik Asooja, Jorge Gracia, Paul Buitelaar, and Philipp Cimiano. Domain adaptation for ontology localization. *Journal of Web Semantics*, 36:23–31, 2016. doi:10.1016/J.WEBSEM.2015.12.001.

64 John Philip McCrae, Christian Chiarcos, Francis Bond, Philipp Cimiano, Thierry Declerck, Gerard de Melo, Jorge Gracia, Sebastian Hellmann, Bettina Klimek, Steven Moran, Petya Osenova, Antonio Pareja-Lora, and Jonathan Pool. The Open Linguistics Working Group: Developing the Linguistic Linked Open Data Cloud. In *Proceedings of the 10th Language Resources and Evaluation Conference (LREC 2016)*, pages 2435–2441, Paris, France, 2016. URL: http://www.lrec-conf.org/proceedings/lrec2016/pdf/851_Paper.pdf.

65 George A. Miller. Wordnet: A lexical database for english. *Commun. ACM*, 38(11):39–41, 1995. doi:10.1145/219717.219748.

66 Marc Miquel-Ribé and David Laniado. Wikipedia culture gap: quantifying content imbalances across 40 language editions. *Frontiers in Physics*, 6:54, 2018. doi:10.3389/fphy.2018.00054.

67 Carmen Moors, Ilana Wilken, Karen Calteaux, and Tebogo Gumede. Human language technology audit 2018: Analysing the development trends in resource availability in all south african languages. In *Proceedings of the Annual Conference of the South African Institute of Computer Scientists and Information Technologists*, SAICSIT '18, pages 296–304, New York, NY, USA, 2018. Association for Computing Machinery. doi:10.1145/3278681.3278716.

68 Diego Moussallem, Axel-Cyrille Ngonga Ngomo, Paul Buitelaar, and Mihael Arcan. Utilizing knowledge graphs for neural machine translation augmentation. In *Proceedings of the 10th international conference on knowledge capture*, pages 139–146, 2019. doi:10.1145/3360901.3364423.

69 Diego Moussallem, Tommaso Soru, and Axel-Cyrille Ngonga Ngomo. THOTH: neural translation and enrichment of knowledge graphs. In Chiara Ghidini, Olaf Hartig, Maria Maleshkova, Vojtech Svátek, Isabel F. Cruz, Aidan Hogan, Jie Song, Maxime Lefrançois, and Fabien Gandon, editors, *The Semantic Web - ISWC 2019 - 18th International Semantic Web Conference, Auckland, New Zealand, October 26-30, 2019, Proceedings, Part I*, volume 11778 of *Lecture Notes in Computer Science*, pages 505–522. Springer, 2019. doi:10.1007/978-3-030-30793-6_29.

70 Diego Moussallem, Matthias Wauer, and Axel-Cyrille Ngonga Ngomo. Machine Translation using Semantic Web Technologies: A Survey. *Journal of Web Semantics*, 51:1–19, 2018. doi:10.1016/J.WEBSEM.2018.07.001.

71 Arijit Nag, Bidisha Samanta, Animesh Mukherjee, Niloy Ganguly, and Soumen Chakrabarti. A Data Bootstrapping Recipe for Low Resource Multilingual Relation Classification. *arXiv preprint*, 2021. doi:10.48550/arXiv.2110.09570.

72 Roberto Navigli and Simone Paolo Ponzetto. Babelnet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial Intelligence*, 193:217–250, 2012. doi:10.1016/J.ARTINT.2012.07.001.

73 Finn Årup Nielsen. Lexemes in wikidata: 2020 status. In Maxim Ionov, John P. McCrae, Christian Chiarcos, Thierry Declerck, Julia Bosque-Gil, and Jorge Gracia, editors, *Proceedings of the 7th Workshop on Linked Data in Linguistics, LDL@LREC 2020, Marseille, France, May 2020*, pages 82–86. European Language Resources Association, 2020. URL: https://aclanthology.org/2020.ldl-1.12/.

74 Aleksandr Perevalov, Andreas Both, Dennis Diefenbach, and Axel-Cyrille Ngonga Ngomo. Can machine translation be a reasonable alternative for multilingual question answering systems over knowledge graphs? In *Proceedings of the ACM Web Conference 2022*, pages 977–986, 2022. doi:10.1145/3485447.3511940.

75 Aleksandr Perevalov, Andreas Both, and Axel-Cyrille Ngonga Ngomo. Multilingual question answering systems for knowledge graphs – A survey. *Semantic Web*, 2023. URL: https://www.semantic-web-journal.net/system/files/swj3417.pdf.

76 Aleksandr Perevalov, Axel-Cyrille Ngonga Ngomo, and Andreas Both. Enhancing the accessibility of knowledge graph question answering systems through multilingualization. In *2022 IEEE 16th International Conference on Semantic Computing (ICSC)*, pages 251–256. IEEE, 2022. doi:10.1109/ICSC52841.2022.00048.

77 Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick S. H. Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander H. Miller. Language models as knowledge bases? In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan, editors, *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 2463–2473. Association for Computational Linguistics, 2019. doi:10.18653/V1/D19-1250.

78 A. Phillips and M. Davis. Tags for Identifying Languages, sep 2009. URL: https://www.rfc-editor.org/info/bcp47.

79 Alessandro Piscopo, Christopher Phethean, and Elena Simperl. Wikidatians are born: Paths to full participation in a collaborative structured knowledge base. In Tung Bui, editor, *50th Hawaii International Conference on System Sciences, HICSS 2017, Hilton Waikoloa Village, Hawaii, USA, January 4-7, 2017*, pages 1–10. ScholarSpace / AIS Electronic Library (AISeL), 2017. doi:10.24251/HICSS.2017.527.

80 S. Ranathunga, E-S. A. Lee, M.P. Skenduli, R. Shekar, M. Alam, and R. Kaur. Neural machine translation for low-resource languages: A survey. *CoRR*, abs/2106.15115, 2021. doi:10.48550/arXiv.2106.15115.

81 Aarne Ranta. *Multilingual Text Generation for Abstract Wikipedia in Grammatical Framework: Prospects and Challenges*, pages 125–149. Springer International Publishing, Cham, 2023. doi:10.1007/978-3-031-21780-7_6.

82 Georg Rehm and Andy Way, editors. *European Language Equality: A Strategic Agenda for Digital Language Equality*. Cognitive Technologies. Springer, 2023. doi:10.1007/978-3-031-28819-7.

83 Bhavyajeet Singh, Pavan Kandru, Anubhav Sharma, and Vasudeva Varma. Massively Multilingual Language Models for Cross Lingual Fact Extraction from Low Resource Indian Languages. *arXiv preprint*, 2023. doi:10.48550/arXiv.2302.04790.

84 Yuan Sun, Yan Zhuang, Sisi Liu, and Xiaobing Zhao. Low-resource language question generation based on key sentence and knowledge graph. *Available at SSRN 4560896*, 2023. doi:10.2139/ssrn.4560896.

85 Thomas Pellissier Tanon and Lucie-Aimée Kaffee. Property label stability in wikidata: Evolution and convergence of schemas in collaborative knowledge bases. In Pierre-Antoine Champin, Fabien Gandon, Mounia Lalmas, and Panagiotis G. Ipeirotis, editors, *Companion of the The Web Conference 2018 on The Web Conference 2018, WWW 2018, Lyon , France, April 23-27, 2018*, pages 1801–1803. ACM, 2018. doi:10.1145/3184558.3191643.

86 Vinh Tong, Dat Quoc Nguyen, Trung Thanh Huynh, Tam Thanh Nguyen, Quoc Viet Hung Nguyen, and Mathias Niepert. Joint multilingual knowledge graph completion and alignment. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang, editors, *Findings of the Association for Computational Linguistics: EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 4646–4658. Association for Computational Linguistics, 2022. doi:10.18653/V1/2022.FINDINGS-EMNLP.341.

87 Y. Tsvetkov. Opportunities and challenges in working with low-resource languages, 2017. URL: https://www.cs.cmu.edu/ ytsvetko/jsalt-part1.pdf.

88 Preeti Vats, Nonita Sharma, and Deepak Kumar Sharma. Hkg: A novel approach for low resource indic languages to automatic knowledge graph construction. *ACM Transactions on Asian and Low-Resource Language Information Processing*, 2023. doi:10.1145/3611306.

89 Denny Vrandecic. Building a multilingual wikipedia. *Communications of the ACM*, 64(4):38–41, 2021. doi:10.1145/3425778.

90 Amrapali Zaveri, Anisa Rula, Andrea Maurino, Ricardo Pietrobon, Jens Lehmann, and Sören Auer. Quality assessment for linked data: A survey. *Semantic Web*, 7(1):63–93, 2016. doi:10.3233/SW-150175.

91 Wenxuan Zhou, Fangyu Liu, Ivan Vulic, Nigel Collier, and Muhao Chen. Prix-LM: Pretraining for Multilingual Knowledge Base Construction. *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 5412–5424, 2022. doi:10.18653/V1/2022.ACL-LONG.371.

92 Yucheng Zhou, Xiubo Geng, Tao Shen, Wenqiang Zhang, and Daxin Jiang. Improving zero-shot cross-lingual transfer for multilingual question answering over knowledge graph. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5822–5834, 2021. doi:10.18653/V1/2021.NAACL-MAIN.465.