Cross-lingual link discovery for under-resourced languages

Michael Rosner, Sina Ahmadi, Elena-Simona Apostol, Julia Bosque-Gil, Christian Chiarcos, Milan Dojchinovski, Katerina Gkirtzou, Jorge Gracia, Dagmar Gromann, Chaya Liebeskind, Giedrė Valūnaitė Oleškevičienė, Gilles Sérasset, Ciprian-Octavian Truic

Angaben zur Veröffentlichung / Publication details:

Rosner, Michael, Sina Ahmadi, Elena-Simona Apostol, Julia Bosque-Gil, Christian Chiarcos, Milan Dojchinovski, Katerina Gkirtzou, et al. 2022. "Cross-lingual link discovery for under-resourced languages." In *13th Language Resources and Evaluation Conference (LREC 2022), 20-25 June 2022, Marseille, France*, edited by Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, et al., 181–92. Stroudsburg, PA: Association for Computational Linguistics.

Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), pages 181–192
Marseille, 20-25 June 2022

Cross-Lingual Link Discovery for Under-Resourced Languages

Michael Rosner[1], Sina Ahmadi[2], Elena-Simona Apostol[3,12], Julia Bosque-Gil[4], Christian Chiarcos[5], Milan Dojchinovski[6], Katerina Gkirtzou[6], Jorge Gracia[4], Dagmar Gromann[10], Chaya Liebeskind[11], Giedrė Valūnaitė Oleškevičienė[8], Gilles Sérasset[9], Ciprian-Octavian Truică[3,12]

[1]University of Malta
[2]NUI Galway, Ireland
[3]Faculty of Automatic Control and Computers, University Politehnica of Bucharest, Romania
[4]Aragon Institute of Engineering Research (I3A), University of Zaragoza, Spain
[5]Applied Computational Linguistics, Goethe University, Frankfurt, Germany
[6]Faculty of Information Technologies, CTU in Prague, Czech Republic
[7]Institute of Language and Speech Processing, Athena Research Center, Greece
[8]Institute of Humanities, Mykolas Romeris University, Lithuania
[9]Univ. Grenoble Alpes, CNRS, Grenoble INP, LIG, Grenoble, France
[10]University of Vienna, Austria
[11]Jerusalem College of Technology, Israel
[12]Department of Information Technology, Uppsala University, Sweden
Corresponding author: mike.rosner@um.edu.mt

Abstract

In this paper, we provide an overview of current technologies for cross-lingual link discovery, and we discuss challenges, experiences and prospects of their application to under-resourced languages. We first introduce the goals of cross-lingual linking and associated technologies, and in particular, the role that the Linked Data paradigm (Bizer et al., 2011) applied to language data can play in this context. We define under-resourced languages with a specific focus on languages actively used on the internet, i.e., languages with a digitally versatile speaker community, but limited support in terms of language technology. We argue that languages for which considerable amounts of textual data and (at least) a bilingual word list are available, techniques for cross-lingual linking can be readily applied, and that these enable the implementation of downstream applications for under-resourced languages via the localisation and adaptation of existing technologies and resources.

Keywords: linguistic data, linguistic linked open data, linking, link discovery

1. Introduction

As reported by the Nexus Linguarum[1] COST Action in a recent policy brief (Bosque-Gil et al., 2021), the possibilities for language communities to exploit new technological advances depends crucially on the availability and abundance of richly annotated linguistic data, which is by definition scarce in under-resourced languages.

The lack of annotated data is damaging in two ways: (i) it hinders the use of advanced data-intensive techniques such as deep learning, multilingual embedding etc, to create a minimum set of *basic* NLP technologies and (ii) it hinders the use of those same techniques for automated enrichment of resources in order to develop *more advanced* NLP technologies.

This leaves under-resourced languages in a permanent state of limbo unless something is done to break the vicious circle linking the development of tools and annotated data.

[1]https://nexuslinguarum.eu/

1.1. The Multilingual Web of Data

A possible route out of the circle is proposed by Gracia et al. (2012a) who envision a Multilingual Web of Data (MWD) which would support universal access to content in any natural language by relying on explicit mechanisms to exploit and reconcile multilingual data automatically. This would also be a step towards bridging the gap between language-independent semantic content and language- and culture-specific information needs, in addition to mitigating the bias in the Semantic Web towards English and other resource-rich languages.

A relatively recent, but continuously growing trend in this field is the increased adoption of Linked Data (Bizer et al., 2011) for representing language resources, a technology that was originally designed to create synergies between data sources in the Web of Data. Linguistic Linked Open Data (LLOD), introduced in 2011 by Chiarcos et al. (2011b), has been the focus of intense research (for a survey see Khan et al., accepted), and subsequent applications in different industries (Aguado-de Cea et al., 2016b; Gracia et al., 2020; Wetzel, 2021; Saurí and Grosse, 2021), including at

present four industry-led pilot projects that address various facets of cross-language transfer or domain adaptation.

Gracia et al. (2012a) propose an architecture and a roadmap for the implementation of the MWD and enumerate a series of prerequisites including (i) Linked Data as the core of the architecture, (ii) the creation of a linguistic layer that describes the original LD data sources without modifying them, and (iii) an architecture that hinges on two main components, namely: a layer of multilingual information (composed from the multilingual linguistic information and multilingual mappings across the data), and a set of services and models on top of this layer. These services and models would help to represent, create and provide support to the access to multilingual information available in the former layer of the LD infrastructure. Among these services the authors envisage services for generating multilingual linked data, services for translation and ontology localization, services for cross-lingual linkage, and services for cross-lingual access.

Cross-lingual interlinking is a complex multi-stage process as described further in Gracia et al. (2012b), and this paper forms part of a wider investigation into the role of cross-lingual interlinking towards the realisation of MWD. In this paper we focus in particular on the stage of cross-lingual link discovery. This involves two subproblems - deciding *what* resources should be linked together and addressing the additional complexity arising from the presence of more than one language.

Of particular importance for cross-lingual and multilingual technologies in this context is the OntoLex vocabulary,[2] widely used for publishing lexical resources on the web, and specifically designed to facilitate linking between dictionaries and knowledge graphs (ontologies), but also applied in a broad number of applications in digital lexicography, language technology and the language sciences. One consequence of its popularity is that the majority of data sets currently available from the Linguistic Linked Open Data cloud.[3] are lexical resources.

With the OntoLex vocabulary published in 2016, and its subsequently increasing adaptation for a large number of use cases, ranging from lexicography (Bosque-Gil et al., 2019) to language documentation (Chiarcos et al., 2017b), linking across different vocabularies has since then not only become much easier (as the data is less heterogeneous and more easily accessible on the web), but also more inclusive to under-resourced languages, as on this basis bilingual dictionaries for hundreds of languages have been made available in machine-readable form.

These technologies represent a basis for developing a truly multilingual, and linguistically inclusive web with minimal technological barriers for speakers of both well-resourced and under-resourced languages. Linked Data is a core technology to achieve cross-lingual linking, with particularly high potential in scenarios in which small, scattered and heterogeneous pieces of information need to be integrated, and its application to *research* on under-resourced languages has been discussed for more than a decade (Poornima and Good, 2010).

1.2. Classifying Under-resourced Languages

(Moran and Chiarcos, 2020) distinguish four degrees of under-resourcedness defined by representative linguistic resources they are lacking:

1. Lack of language data – a general lack of language documentation and description (no substantial grammars, dictionaries, or corpora).[4]
2. Lack of accessible language data – resources exist but not in a form that can be easily processed, e.g., because they are not available in digital form or accessing them requires legacy hardware or proprietary software.[5]
3. Lack of language technological support – there is accessible language data, e.g., texts from the internet or social media, but there are no or insufficient NLP tools, linguistic annotations, lexical resources,[6]

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

[3] https://linguistic-lod.org/

[4] Typical category 1 languages are languages that are endangered or spoken in remote areas such as, for example, Jarawa (ISO 639-3 anq), an Ongan language spoken on the Andaman Islands, less than 300 speakers. Available language material includes word lists as part of linguistic treatises on selected aspects of grammar but no extant texts. The language is not used on the internet. A possible sister language, Sentinelese (ISO 639-3 std) is fully undocumented but only *suspected* to be related for geographical reasons and ethnographic parallels.

[5] Khinalug (ISO 639-3 kjj), 3,000 speakers, spoken in the Caucasus – available language material (in print, not digital) consists of three grammars with word lists and glosses, one dictionary (unpublished, so far), a partial Bible translation (unpublished) and a few brochures with poems and stories used for teaching the language to the next generation (only available locally). The language does not have a standard orthography and is not used on the internet.

[6] This is the situation for the majority of languages used over the internet, including widely used languages such as Hausa (ISO 639-3 hau), the 2nd most used language of Africa. We are not aware of publicly available Hausa corpora with linguistic annotations, but the language is widely used on the internet, and because of the abundant amounts of data, there have been academic experiments with Hausa corpora, e.g., (Chiarcos et al., 2011a), as well as applications of machine translation (e.g., in Google translate). However, a systematic development of dedicated NLP tools and the necessary pre-requisites seems to have begun only about a year ago (https://github.com/hausanlp/hausanlp). The primary electronic dictionary for Hausa (http://maguzawa.dyndns.ws/) dates from the 1930s (Bargery,

4. Limited interoperability of available data and tools – there are annotated corpora, tools and lexical resources, but they are not designed to work together, so that the language does not have a complete and consistent NLP stack. This is the typical situation of smaller national languages and roughly corresponds to the status of English about 25 years ago.[7]

Categories 1 and 2 apply to languages that lack usable digital data, e.g., historical languages, small speaker communities in developed countries, speaker communities in remote areas or speaker communities marginalized for political reasons, and these are what Moran and Chiarcos (2020) have been focusing on as they pertain the most elementary needs of researchers and speakers. Category 4, on the other hand, has been the original motivation for the application of the Linked Open Data paradigm to language technology and the language sciences, so that this aspect is exceptionally well covered, e.g., in the recent textbook by Cimiano et al. (2020a). Category 3, however, has experienced less coverage, and we address this gap by discussing the role that technologies for cross-lingual link discovery can play in this context. We specifically focus on languages that are being actively used for communicating over the web (so, textual data, and, most likely, grammatical descriptions and word lists are available) and the needs of their speakers to participate in the flow of information and goods over the internet.

1.3. Dealing with Category 3 Languages

This set of languages has been given various labels in the literature. Perhaps the oldest is "low-density languages" (Jones and Havrilla 1998). The terms "medium- density" and "lower-density languages" have also been coined (e.g., Maxwell and Hughes 2006). The latter term specifically refers to "the amount of computational resources available, rather than the number of speakers any given language might have" (Maxwell and Hughes 2006; Meyers et al. 2007). The amount of accessible data, regardless of language-speaker quantities, is the theme that binds these various terms together.

Moran and Chiarcos (2020) provide a recent overview of the application of Linguistic Linked Open Data technology for creating and using datasets for under-resourced languages, but with a focus on applications in linguistics and the humanities. Most notably, their understanding of under-resourced languages particularly pertains to challenges in language documentation and linguistic typology, i.e., scenarios where not even substantial amounts of (digital or other) text are available, but merely field notes, linguistic treatises and word lists.

This paper complements the discussion of under-resourced languages and linking technologies with a different angle, by focusing on languages that are actively used, e.g., on the web and in social media, but which have limited (or no) technological support in terms of language technology or information technology in general. In this scenario, we can expect that textual data and dictionaries (or word lists) are available, but neither extensive language resources nor richly and deeply annotated corpora or tools to produce such annotations automatically. The question our paper tries to answer is how cross-lingual linking techniques can be used to facilitate the development of language technology and language resources for under-resourced languages.

While we focus on under-resourced languages for which there is limited language technology, we make what we believe to be reasonable minimal assumptions concerning availability of data, namely that the language has (i) at least rudimentary lexical resources (say, word lists), but not more mature resources such as WordNets or rich ontologies, and (ii) that we have digital text, but no significant amounts of annotated data from which to train state-of-the-art NLP tools.

The fundamental challenge here is to exploit languages with richer resources or more developed language technologies to benefit the under-resourced language from their respective solutions. In this paper we argue that the discovery, creation and deployment of appropriate cross-lingual links is of key importance in addressing that challenge.

2. Cross-lingual Link Discovery

Cross-lingual links provide the means to align material in one language to material in another and are thus the building blocks for different multilingual resources. These links are a way of storing information that adds value to all resources linked and that could be difficult to find otherwise (Chiarcos et al., 2013). Once established, links among resources and

1934) and uses an outdated orthography. Moreover, we are not aware of any tools or dictionaries that support *both* (Latin-based) Boko and (Arabic-based) Ajami orthographies of Hausa so that language technology coverage for the language is partial, at best.

[7] Georgian (ISO 639-3 kat) can be considered a Category 4 language. There are substantial web corpora (e.g., `https://wortschatz.uni-leipzig.de/en/download/Georgian`), dictionaries (e.g., `https://github.com/acoli-repo/acoli-dicts/blob/master/stable/panlex/biling-tsv/panlex-20191001-csv-tsv/ka.zip`), and a national corpus with dedicated NLP tools that are available as web services `http://gnc.gov.ge/gnc`, but also a number of independent efforts which are not coordinated with each other, e.g., a syntactically annotated parallel corpus (Kapanadze, 2014, GRUG), and a number of annotation tools developed at the Ilia State University, Georgia, e.g., the QartNLP lemmatizer `https://qartnlp.iliauni.edu.ge/lemma`. With the latter, the GNC and GRUG morphologies, at least three independent finite state morphologies for Georgian seem to be in existence.

languages reduce the effort needed to create and enrich future resources. Resources aligned thus have been shown to improve word, knowledge and domain coverage. They serve in the creation of new multilingual lexical resources such as ConceptNet (Speer et al., 2017), Yago (Suchanek et al., 2007) and BabelNet (Navigli and Ponzetto, 2012a). In addition, they can improve performance of NLP tasks such as word sense disambiguation (Navigli and Ponzetto, 2012c), semantic role tagging (Xue and Palmer, 2004) and semantic relations extraction (Swier and Stevenson, 2005).

In general the process of creating links from existing language resources essentially involves two steps: (i) discovery of which data items in different RDF datasets to link with respect to a given relation R, and (ii) linking them together by creating a new RDF triple using e.g. `owl:sameAs`.

The link discovery problem with respect to a relation R expressed by the link is defined thus: given two sets of resources S and T, find all pairs $(s, t) \in S \times T$ such that $R(s, t)$ holds (Nentwig et al., 2017). Useful analyses of the link discovery process are provided in surveys by these authors as well an earlier one by Ferrara and Nikolov (2011).

Both surveys make it clear that matching operation which decides whether a pair of data items satisfy the R is central, and may be set up in different ways according to the kind of items linked, the nature of the relation R etc. For example, Cimiano et al. (2020a), suggest a four-way classification of link discovery techniques: *terminological* (use of string-based comparison methods and linguistic methods to compare text); *structural* (exploit the organisation and internal structure of dataset elements); *extensional* (employ the so-called extension of classes in terms of individual elements sharing the class in the set), and *semantic* (based on the model-theoretic semantics of RDF and OWL).

The two most extensively used frameworks for link discovery are SILK (Volz et al., 2009) and LIMES (Ngonga Ngomo and Auer, 2011) both of which offer a user interfaces for customising the discovery process and manipulating data and links. LIMES in addition focuses on the incorporation of machine learning algorithms and optimisation of the search process.

Unfortunately, few such frameworks have cross-lingual capabilities. One example is Lesnikova (2013), where the authors proposes a language-oriented approach to align RDF graphs based on natural language terms present within the graphs. In this approach, a comparison is carried out between the language data of each URI in two datasets. To this end, the RDF graph is converted into documents, called virtual documents. The usage of virtual documents is motivated by the idea that identical entities across two graphs can be compared independently from their structures. As a result, the most similar representations are identified using machine translation techniques and a vector space model. Finally, the aligned triples are created with the `owl:sameAs` property.

A comprehensive overview of link discovery and representation which also includes consideration of the cross-lingual case is given by Cimiano et al. (2020b).

3. Creating Resources with Cross-lingual Links

Above we have suggested that cross-lingual link discovery is of key importance in helping to create new enriched resources which serve for development of NLP technologies for under-resourced languages. In this section, we shed light on some of the major tasks where cross-lingual link discovery has indeed played such a role with respect to well-resourced languages, providing a concise but essential survey on previous methods and approaches. Section 4 then describes their application to low-resource languages.

3.1. Resource Creation and Enrichment

Cross-lingual link discovery has related to both the creation of new resources from existing ones, and to the enrichment of existing resources by adding more data to increase diversity and multilingualism.

In this regard, Sánchez-Rada and Iglesias (2016) propose a LD approach to represent emotion with a focus on lexical resources and emotion analysis services. The ontology that models emotion analysis, called Onyx, comes with a semantic vocabulary of emotions that is integrated with a few other well-known vocabularies, such as Lemon[8], NIF[9] and the Provenance Ontology[10]. Moreover, Onyx provides formalisms to align lexical entries with external resources such as WordNet (Miller, 1995) and DBpedia[11]. Therefore, this ontology increases interoperability across resources in different languages by bridging LD with semantic and emotion analysis.

Caracciolo et al. (2012) describe the creation and maintenance of the AGROVOC multilingual thesaurus in LD. This thesaurus covers areas of interest to the Food and Agriculture Organisation (FAO) and is aligned with other multilingual knowledge organisation systems related to agriculture, using the SKOS properties exact match and close match. Alignments are automatically produced using a custom-designed tool based on string similarity matching algorithms. The candidate mappings are then validated by a domain expert who considered `skos:exactMatch` between validated entities.

Interlinear glossed text (IGT) is a popular notation used in various fields of linguistics and provides syntactic and semantic annotations that allow the reader to

[8] `https://lemon-model.net/`

[9] `https://persistence.uni-leipzig.org/nlp2rdf/ontologies/nif-core/nif-core.html`

[10] `https://www.w3.org/TR/prov-o/`

[11] `https://www.dbpedia.org/`

follow the relationship between a source text and its translation. (Chiarcos et al., 2017a) propose a representation of IGT data in RDF, along with converters from two popular IGT frameworks and an automatic linking procedure with dictionaries, such as DBnary (Gilles Sérasset, 2012 2022), (Sérasset, 2015) multilingual resource. The proposed RDF representation uses a shallow RDFS data model, isomorphic to the original data structures which does not provide rich semantics. However, it does guarantee that representations are transparent, familiar to their user community, and easily linked to external resources.

BabelNet (Navigli and Ponzetto, 2012b), a well-known multilingual semantic network that integrates WordNet[12], a lexicographic resource, and Wikipedia[13], an encyclopedic knowledge resource, has been converted into a LD representation, using *lemon* (Ehrmann et al., 2014). The result is an interlinked multilingual lexical resource accessible as LOD which can be used not only to enrich existing datasets with linguistic information, but also to support the process of mapping datasets across languages[14].

3.2. Bilingual Lexicon Induction

Bilingual Lexicon Induction (BLI), also known as translation inference, is the task of inducing new translation pairs based on monolingual, bilingual or multilingual dictionaries or lists of words. Thus, one way to create a bilingual lexicon between languages X and Z is by triangulation, whereby we make use of a pair of existing bilingual lexicons that share a common language Y (given lexicons X-Y and Y-Z, bootstrap X-Z) or a common word-sense representation.

A bilingual dictionary associates terms in one language to equivalent terms in a second language by representing cross-lingual links in an efficient way. The creation of any such resource requires links to be explicitly represented, and this can only take place after a discovery phase that identifies the term or terms in a target language that are semantically equivalent to a given source term. A range of discovery methods have been proposed e.g. based on machine learning (Donandt et al., 2017), graph-based algorithms (Villegas et al., 2016; Torregrosa et al., 2019), or the use of an intermediate pivot language (Tanaka and Umemura, 1994).

The need to evaluate BLI approaches has motivated the TIAD (Translation Inference Across Dictionaries) shared task[15] which started in 2017 and is aimed at *exploring methods and techniques for automatically generating new bilingual (and multilingual) dictionaries from existing ones in the context of a coherent experiment framework that enables reliable validation of results and solid comparison of the processes used*.

There are several ways in which LD can play a crucial role in BLI. First, by providing standard representation mechanisms for lexicons and translations between them. Second, by providing access and querying mechanisms to graphs of bilingual dictionaries on the Web. Finally, the graph-based techniques for BLI encounter in LD graphs a natural application scenario, given the graph nature of RDF. This is illustrated by the fact that TIAD is using the Apertium RDF graph (Gracia et al., 2018) as basis of its recent campaigns (Kernerman et al., 2020; Gracia et al., 2019).

An approach to support BLI from a modelling point of view is exemplified in Aguado-de Cea et al. (2016a) which describes how bilingual connections across a set of dictionary resources for different languages, such as K Dictionaries[16], can be transformed into a linked representation using the Ontolex-Lemon Vartrans module (Bosque-Gil et al., 2017).

A way to achieve BLI is by searching whether monolingual links already exist from source term w_s and target term w_t to a shared sense inventory. The search process minimally involves computation of the following relation:

$$\{< w_s, w_t > | sense(w_s, \sigma) \wedge sense(w_t, \sigma)\}$$

that is, pairs of source and target words sharing a sense σ where $sense(w, s)$ embodies the relation between a word w and a word sense s. Clearly this computation involves two monolingual sense queries and an intersection operation. This is actually the way in which links between Apertium RDF (Gracia et al., 2018) and the LD version of BabelNet (Ehrmann et al., 2014) (acting as sense inventory) were established.

Similarly, Fang et al. (2016) describe the process of creating a Chinese lexicon enriched with links to DBpedia (using the equivalence relations found in DBpedia) and BabelNet (using BabelNet category labels to disambiguate and link synsets).

3.3. Lexical Alignment

Conversion of an existing lexical resource to an established LOD format facilitates alignment with other resources on the LLOD cloud, which not only enriches the original resource, but also the resources to which the links are made.

A concrete example is lemonUBY (Eckle-Kohler et al., 2015), created by the conversion of UBY (Eckle-Kohler et al., 2012), a large repository of lexical resources, to lemon, thus facilitating the production of links to the OLiA reference model, and in turn to universal linguistic terminologies such as GOLD and ISOcat. Provided that annotations are correct, links between lemonUBY and other lexical resources in the LLOD and can sometimes be discovered automatically based on shared lemma and POS information.

[12] https://wordnet.princeton.edu/

[13] https://www.wikipedia.org/

[14] The last available version of BabelNet as LLOD is 3.6. Later updates do not contain updates of the LD version

[15] https://tiad2021.unizar.es/

[16] https://www.lexicala.com/k-dictionaries

Another example of resource creation though lexical alignment is LIDIOMS Moussallem et al. (2018), a multilingual resource in LD format containing multiword expressions in five languages (English, German, Italian, Portuguese, Russian). String similarities were the basis for the discovery of links to other existing LD resources DBnary (Sérasset, 2015) and BabelNet (Navigli and Ponzetto, 2012b) with LIMES algorithms (Ngonga Ngomo, 2012) for the former, through the `rdfs:label` property using trigram similarity to which an acceptance threshold of 0.85 was set. For links to BabelNet, a manual comparison was made between the `skos:definition` property from LIDIOMS and the `bn-lemon:definition` property from BabelNet.

4. Application to Under-Resourced Languages

With the technologies described in the last section, numerous novel functionalities can be enabled for under-resourced language. Below, we focus on category 3 languages - those for which, minimally, we have at least substantial amounts of digital text available and a word list.

4.1. Methods

On this basis, three general methods can be envisaged as a result of linking: (i) the induction of additional lexical resources, (ii) the enrichment of existing knowledge graphs and ontologies with labels in the under-resourced language and (iii) the dynamic multilingual querying of knowledge bases by means of lexical databases and the federation features of SPARQL.

4.1.1. Lexicon Induction

We have seen how LD-based BLI allows the discovery of new translation relations between initially disconnected language pairs. This can bring direct benefits to under-resourced languages, with the potential of creating a number of new multilingual/bilingual resources around a given language and thus enabling direct and indirect connections with other resources in other languages.

For instance, the Apertium RDF graph contains lexicons and translations of a number of minority languages (e.g., Aragonese, Occitan, Esperanto, Maltese, etc.) which are now part of a single unified graph and have the potential, through BLI methods, to be further enriched with translations into more languages. The net effect, then, is to reduce the level of under-resourcedness of the language in question.

It is important to emphasize that the possibility to link into a graph of resources is not in itself a solution to the problem of under-resourced languages. However, it provides a basic level of infrastructure that opens the door to a wide variety of (as yet undiscovered) methods for searching, exploring and discovering relations between data elements in under-resourced languages, which are relatively isolated, and those in well-resourced languages, which are not.

4.1.2. Multilingual enrichment

Unlike resource creation creation and enrichment as introduced above, a specific application scenario for under-resourced languages is the lexicalization of a foreign language knowledge graph into those languages. Minimally, this means augmenting an existing resource with new labels in the under-resourced language. With such target language labels, any knowledge graph can be queried and processed for application with data from the under-resourced language.

As an example, Araúz et al. (2011) use multilingual labels as a simple but powerful method for conceptual matching. In particular, they link their EcoLexicon, a multilingual terminological knowledge base on the environment, with DBpedia (Lehmann et al., 2012), GeoNames[17], and GEMET, the GEneral Multilingual Environmental Thesaurus[18]. The discovered links are represented using *owl:sameAs*. Their method takes all the English variants of a term expanded with equivalences in other languages and explores coincidences with the multilingual labels of the target term. There is a problem, though, if polysemy occurs at a cross-linguistic level. In that case, they add category membership information to the linking algorithm (e.g., to indicate domains such as "geography" or "oceanography").

This methodology can be trivially applied to any language for which a machine-readable bilingual word list is available, so that novel, localized knowledge bases can be bootstrapped with minimal effort. Such linking tasks are particularly easy to perform if word lists (or bilingual dictionaries) are available in a machine-readable form that is compatible with the knowledge graph under consideration, i.e., by using web standards such as RDF and update (delete/insert) and construct operators in SPARQL(Harris et al., 2013) over RDF-encoded data. Indeed, this has been one motivation for developing the OntoLex vocabulary and for linking OntoLex-encoded dictionaries with a knowledge graph, where OntoLex provides an RDF vocabulary for lexical resources on the web (and indeed, bilingual dictionaries for many under-resourced languages are being provided in OntoLex, e.g., (Westphal et al., 2015; Chiarcos et al., 2020; Abromeit et al., 2016)).

Moreover, the application of SPARQL is not limited to compiling novel resources. In fact, the concept of federation (Buil-Aranda et al., 2013) is a fundamental design principle of SPARQL, so that *distributed* resources can also be processed on-the-fly. This is illustrated in a small show-case for cross-lingual querying.

[17] `https://www.geonames.org/ontology/`

[18] `https://www.eionet.europa.eu/gemet/en/themes/`

```
SELECT DISTINCT * WHERE {
  ?entry ^dbnary:isTranslationOf
         /dbnary:writtenForm "bagan"@bm.
  ?t dbnary:isTranslationOf ?entry;
     dbnary:targetLanguage lexvo:eng;
     dbnary:writtenForm ?translation.
  SERVICE <https://dbpedia.org/sparql> {
    SELECT DISTINCT * WHERE {
      ?a rdfs:label ?translation.
      ?b a ?a
    }
  }
} LIMIT 100
```

Listing 1: Querying DBpedia from Bambara term "bagan" (animal)

4.1.3. Cross-lingual querying

Thanks to the ever-growing number of language resources available in the LLOD, it becomes possible to easily provide cross-lingual querying services even for under-resourced languages.

Such a service can easily be crafted using any existing SPARQL endpoint offering federated queries. As an example we present a simple scenario where users of an under-resourced language (e.g. Bambara) want to query the DBpedia (Lehmann et al., 2012) ontology for instances of category "bagan"@bm (animal). It must be noted that DBpedia does not incorporate the (still embryonic) Bambara wikipedia language edition.

The service works in two steps. First candidate translations are obtained from the query language to the ontology language; then the ontology is queried using the candidates. In this example, we use DBnary (Sérasset, 2015) as a lexical resource for translations. The Bambara query term is not translated to/from English, hence, the query will make use of a pivot entry to translate from Bambara to English.

Listing 1 gives an example of a single query performing both steps thanks to SPARQL federated queries[19].

This toy example could be elaborated further to make use of any other lexicon available in the LLOD cloud, provided that it is accessible through a SPARQL endpoint. Note that this service comes for "free" for the under-resourced language user who does not have to setup any specific hardware or software. Moreover, as the service relies on online resources its quality will evolve with the resources.

In the remainder of this section, we describe applications of these methods and tools built on top of them and show how these could improve the situation of under-resourced languages, respectively case studies in which resources for under-resourced languages have been created.

[19]The reader may want to try this query online at `https://tinyurl.com/bam-bagan-federated-query`

4.2. Named Entity Recognition

Named entity recognition (NER) is the task of identifying and classifying key objects (entities) in text. An entity is a (real, abstract or imagined) thing that words or a string of words in a text refer to. Traditional NER deals primarily with recognising and categorizing entities (e.g. persons, locations, organisations) but if entities are present in a knowledge graph or an ontology, and a link between the text and this knowledge base is created, the activity is referred to as entity linking.

Entity linking is a key technology because it enables a range of downstream applications based on the fundamental connection between language, objects in the world, and the available knowledge about those objects. It also enables semantic search (based on meaning rather than words). NER and entity linking are of great importance, and resources for NER need to address the problem created by the fact that the same entity can have different names in the same language. The problem is compounded when different languages are involved and clearly this is particularly acute for under-resourced languages.

Various approaches tackle the task of linking entities in multiple languages. Blissett and Ji (2019) encode the orthographic similarity of mentions using bidirectional Gated Recurrent Units architecture and then cluster them using DBSCAN with the objective of maximizing the CEAFm F-score. Experiments using Tigrinya and Oromo show a great improvement over edit distance for linking entities.

LODeXporter (Witte and Sateli, 2018) is a flexible component for the GATE (General Architecture for Text Engineering) framework (Cunningham et al., 2013) used to generate URIs for LOD triples from textual data for automatic Knowledge Base construction. This enables a contribution sentence found in a document together with associated concepts (e.g., named entities) to be described and queried using SPARQL to find all documents that contain the sentence as well as all their mentioned topics. The authors conclude that LODeXporter enables NLP framework users to easily generate knowledge bases in a LOD-compliant format and easily connect with different web vocabularies.

NER use cases exist for several under-resourced languages.

For Romanian, Mitrofan (2017) have bootstrapped a corpus for Biomedical Named Entity Recognition (BioNER), a particularly complex task due to the specialized medical terminology that is hard to correctly identify and where there can be multiple name conventions for the same biomedical concept. The corpus consists of $\approx 300k$ sentences and 7 million tokens. $\approx 40k$ tokens were manually annotated and checked by a medical expert. The rest of the labelled Name Entities were automatically detected using a bootstrapping method using Word2Vec to extract the word embeddings, and then a Partitioned Convolutional Neural Network for classification. This corpus was then added to the first

Romanian medical treebank, SiMoNERo (Mititelu and Mitrofan, 2020).

NER use cases are also successfully applied in under-resourced languages Latvian and Lithuanian. These two Baltic languages display rich morphology and feature high morphological ambiguity together with a relatively free order of constituents in sentences, making NER more difficult. Pinnis (2012) presents TildeNER as an open source used for NER in Latvian and Lithuanian. It was evaluated relying on human annotated gold standard test corpora for Latvian and Lithuanian languages. The use case developed toolkit TildeNER comprises wide configuration possibilities for various NER tasks such as aid in question answering, machine translation, keyword extraction, etc., where different requirements for higher precision could be applied.

Ehrmann et al. (2016) have proposed JRC-Names, a named entity resource rendered as LD data using lemon. The resource offers large-scale data integration, e.g. cross-lingual mapping, and web-based content processing, e.g. entity linking. To link entities with their correct counterpart, JRC-Names offer inter-linking between existing datasets, i.e., DBpedia (Lehmann et al., 2012), New York Times, and Talk-Of-Europe, and uses other controlled vocabularies, i.e., LexInfo (Cimiano et al., 2011), OLiA (Chiarcos and Sukhareva, 2015), and LexVo (De Melo, 2015). Importantly, the lemon representation of the resource retains the lexical sense of the named entities for translation purposes.

4.3. Terminology Extraction

Terminology extraction denotes the identification of single- or multi-word terms in texts and ideally relations between groups of synonymous and equivalent terms. For Lithuanian as a representative under-resourced language, Rokas et al. (2020) achieve comparable results to high-resource term extraction with a small dataset of 1,258 manually annotated terms in the cybersecurity domain by utilising a Bi-LSTM model that performed best when trained with BERT embeddings (Devlin et al., 2018). In order to benefit from multilingual pre-trained language models for link discovery, the relational knowledge inherent in these models needs to be explicated. To this end, (Oliveira, 2021) acquire lexico-semantic relations from a pretrained BERT model for Portuguese by predicting entities in lexico-syntactic patterns, e.g. "um [MASK] é uma parte de X_1" (part-of) where the X_1 is given and [MASK] has to be predicted. Similar to high-resource languages, this approach achieved the highest performance for hypernymy. Interlinking resulting relation instances to existing resources, such as the OpenWordNet-PT, would make them available within the Open Multilingual WordNet resources (Gonçalo Oliveira et al., 2021) facilitating cross-lingual linking. (Wachowiak et al., 2021) propose to combine term and relation extraction from monolingual text in a pipeline approach by fine-tuning two separate instances of XLM-R (Conneau et al., 2020). While trained on English and German, the approach has been evaluated on other languages benefiting from the underlying multilingual model, including Romanian and Portuguese. Predicted relations build on a pre-specified typology, including generic, partitive, spatial, origination relations, which eases their alignment across languages.

4.4. Cross-Lingual Embeddings and Translation Inference

Word embeddings are dense real-valued vector representations of words that are closer in vector space if their meaning is similar. In under-resourced languages, the absence of large text corpora for training word embeddings represents a challenge. Cross-lingual word embeddings have been trained by connecting monolingual corpora in two different languages by means of bilingual dictionaries (see (Ruder et al., 2019) for an overview). Adams et al. (2017) train cross-lingual word embeddings for under-resourced languages by showing their ability to perform well even with scarce data on the target language side and plentiful data on the source language side. Thereby, cross-lingual embeddings enable model transfer between resource-rich and under-resourced languages in a common vector space. In Hartung et al. (2020) this idea is developed further in the form of cross-lingual projection, where source and target language are jointly trained utilising bilingual lexicons and only for the former are annotated data required for training a sentiment analysis model. The authors specifically address the application potential of language resources in the LLOD cloud in various multilingual NLP tasks, which has also been practically shown in e.g. (Gracia et al., 2020; Allgaier et al., 2021). For instance, Gracia et al. (2020) gather LLOD translations, specifically from Apertium RDF, to enable transfer of sentiment knowledge from one language to another without the need to retrain the model.

For the task of translation inference, that is, learning new bilingual dictionaries from existing ones, Donandt and Chiarcos (2019) create sense embeddings for the source language building on OntoLex sense annotations in Apertium and predict word embeddings in the target language on this basis. Towards the same end, Lanau-Coronas and Gracia (2020) build on combining graph exploration and cross-lingual word embeddings to derive translations from the Apertium RDF graph that are not directly connected. Chakravarthi et al. (2019) train multiple source and target languages in a Neural Machine Translation (NMT) model simultaneously, relying on transliteration of under-resourced languages on the target side, bringing closely-related languages into a single script. To this end, the authors rely on English WordNet senses to obtain contextual data that share semantic information to be translated to the Dravidian languages Tamil, Telugu, and Kannada. Another powerful scenario of utilising NMT for gen-

erating under-resourced language data from and with LLOD is the generation of text from RDF triples. To this end, Kasner and Dušek (2020) fine-tune the pre-trained NMT model mBART (Liu et al., 2020) for the RDF-to-text generation in English and Russian. These approaches present the utility of LLOD resources for a variety of under-resourced NLP tasks, from translation inference and text generation to sentiment analysis.

5. Conclusion

The claim underlying this paper is that the use of the LD paradigm for language resources can make a significant contribution towards solving the problem of under-resourcedness which exists for the vast majority of the world's languages.

We have tried to substantiate this claim from several different angles, starting with some background on how the vision of the MWD can serve to reconcile, on one hand, the essential language dependency of linguistic resources and on the other, the potential language independency, and hence universal applicability, of the LD framework within which these resources exist.

We provide a short survey illustrating how LD techniques have already been successfully harnessed for the creation and enrichment of resources in the areas of bilingual lexicon induction and lexical alignment.

We then described the application of cross-lingual linking technologies to under-resourced languages identifying three promising uses of cross-lingual linking that are particularly relevant: lexicon induction, multilingual enrichment and cross-lingual querying. Finally we described some applications (NER, Terminology Extraction, Translation Inference using Cross-Lingual Embeddings) that are enabled by the application of cross-lingual linking or language resources created on that basis.

Overall, we find that for category 3 languages, techniques for cross-lingual linking can be readily applied, and that these enable the implementation of downstream applications as well as the localization and adaptation of existing technologies and resources to their respective needs. With the improved availability of lexical data published in accordance with LD principles they can now be directly applied to a large number of under-resourced languages and thus represent a cornerstone for extending tools and knowledge bases.

This paper has mainly been written to demonstrate not only that our claims are feasible, but that they are concrete enough to have been implemented over a range of languages and domains and resources. However, some will rightly object that we have said very little about the quality of results, efficiency of the techniques described, how we go about measuring improvement in the overall resource status of under-resourced languages, etc. This is because as yet an evaluation framework to address these issues is not in place. For true progress to take place, the importance of such a framework cannot be understated and should be developed hand in hand with the methods proposed.

6. Acknowledgements

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) www.cost.eu. This work is also partially supported by the I+D+i project PID2020-113903RB-I00, funded by MCIN/AEI/10.13039/501100011033, 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).
(Abgaz, 2020)

7. Bibliographical References

Abgaz, Y. (2020). Using OntoLex-lemon for representing and interlinking lexicographic collections of Bavarian dialects. In *Proceedings of the 7th Workshop on Linked Data in Linguistics (LDL-2020)*, pages 61–69, Marseille, France, May. European Language Resources Association.

Abromeit, F., Chiarcos, C., Fäth, C., and Ionov, M. (2016). 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*, pages 11–19.

Adams, O., Makarucha, A., Neubig, G., Bird, S., and Cohn, T. (2017). Cross-lingual word embeddings for low-resource language modeling. In *Proceedings of the 15th Conference of the European Chapter of the ACL: Volume 1, Long Papers*, pages 937–947.

Aguado-de Cea, G., Montiel-Ponsoda, E., Kernerman, I., and Ordan, N. (2016a). From dictionaries to cross-lingual lexical resources. *Kernerman Dictionary News*, 24:25 – 31, July.

Aguado-de Cea, G., Montiel-Ponsoda, E., Kernerman, I., and Ordan, N. (2016b). From dictionaries to cross-lingual lexical resources. *Kernerman Dictionary News*, 24:25–31.

Allgaier, K., Veríssimo, S., Tan, S., Orlikowsky, M., and Hartung, M. (2021). Llod-driven bilingual word embeddings rivaling cross-lingual transformers in quality of life concept detection from french online health communities.

Araúz, P. L., Redondo, P. J. M., and Faber, P. (2011). Integrating environment into the linked data cloud. In Werner Pillmann, et al., editors, *Proc. of the 25th International Conference on Informatics for Environmental Protection, EnviroInfo 2011*, pages 370–379. Shaker Verlag, Aachen.

Bargery, G. (1934). *A Hausa-English Dictionary and English-Hausa Vocabulary*. Oxford University Press, London.

Bizer, C., Heath, T., and Berners-Lee, T. (2011). Linked data: The story so far. In *Semantic ser-*

vices, interoperability and web applications: emerging concepts, pages 205–227. IGI global.

Blissett, K. and Ji, H. (2019). Cross-lingual. In *Proceedings of the Second Workshop on Computational Models of Reference, Anaphora and Coreference*, pages 20–25. ACL.

Bosque-Gil, J., Gracia, J., and Montiel-Ponsoda, E. (2017). Towards a Module for Lexicography in OntoLex. In *LDK Workshops*, pages 74–84.

Bosque-Gil, J., Lonke, D., Kernerman, I., and Gracia, J. (2019). Validating the ontolex-lemon lexicography module with k dictionaries' multilingual data. In *Electron. lexicogr. 21st cent., Proc. eLex conf.*

Bosque-Gil, J., Mititelu, V. B., Oliveira, H. G., Ionov, M., Gracia, J., Rychkova, L., Oleskeviciene, G. V., Chiarcos, C., Declerck, T., and Dojchinovsk, M. (2021). 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.

Buil-Aranda, C., Arenas, M., Corcho, O., and Polleres, A. (2013). Federating queries in sparql 1.1: Syntax, semantics and evaluation. *Journal of Web Semantics*, 18(1):1–17.

Caracciolo, C., Stellato, A., Rajbahndari, S., Morshed, A., Johannsen, G., Jaques, Y., and Keizer, J. (2012). Thesaurus maintenance, alignment and publication as linked data: the AGROVOC use case. *International Journal of Metadata, Semantics and Ontologies*, 7(1):65–75.

Chakravarthi, B. R., Arcan, M., and McCrae, J. P. (2019). Wordnet gloss translation for under-resourced languages using multilingual neural machine translation. In *Proceedings of the Second Workshop on Multilingualism at the Intersection of Knowledge Bases and Machine Translation*, pages 1–7.

Chiarcos, C. and Sukhareva, M. (2015). Olia – ontologies of linguistic annotation. *Semantic Web*, 6:379–386.

Chiarcos, C., Fiedler, I., Grubic, M., Hartmann, K., Ritz, J., Schwarz, A., Zeldes, A., and Zimmermann, M. (2011a). Information structure in african languages: corpora and tools. *Language resources and evaluation*, 45(3):361–374.

Chiarcos, C., Hellmann, S., and Nordhoff, S. (2011b). Towards a linguistic linked open data cloud: The open linguistics working group. *Trait. Autom. des Langues*, 52(3):245–275.

Chiarcos, C., McCrae, J., Cimiano, P., and Fellbaum, C. (2013). Towards open data for linguistics: Linguistic linked data. In *New Trends of Research in Ontologies and Lexical Resources*, pages 7–25. Springer.

Chiarcos, C., Ionov, M., Rind-Pawlowski, M., Fäth, C., Schreur, J. W., and Nevskaya, I. (2017a). Llodifying linguistic glosses. In Jorge Gracia, et al., editors, *Language, Data, and Knowledge*, pages 89–103, Cham. Springer International Publishing.

Chiarcos, C., Walther, D., and Ionov, M. (2017b). From language documentation data to llod: A case study in turkic lemon dictionaries. In *LDK Workshops*, pages 22–32.

Chiarcos, C., Fäth, C., and Ionov, M. (2020). The acoli dictionary graph. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 3281–3290.

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

Cimiano, P., Chiarcos, C., McCrae, J. P., and Gracia, J. (2020a). Linguistic linked data in digital humanities. In *Linguistic Linked Data*, pages 229–262. Springer.

Cimiano, P., Chiarcos, C., McCrae, J. P., and Gracia, J., (2020b). *Link Representation and Discovery*, pages 181–196. Springer International Publishing, Cham.

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., and Stoyanov, V. (2020). Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online, July. Association for Computational Linguistics.

Cunningham, H., Tablan, V., Roberts, A., and Bontcheva, K. (2013). Getting more out of biomedical documents with GATE's full lifecycle open source text analytics. *PLoS Computational Biology*, 9(2):e1002854, feb.

De Melo, G. (2015). Lexvo.org: Language-related information for the linguistic linked data cloud. *Semantic Web*, 6(4):393–400.

Devlin, J., Chang, M., Lee, K., and Toutanova, K. (2018). BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Donandt, K. and Chiarcos, C. (2019). Translation inference through multi-lingual word embedding similarity. In *TIAD@ LDK*, pages 42–53.

Donandt, K., Chiarcos, C., and Ionov, M. (2017). Using Machine Learning for Translation Inference Across Dictionaries. In *LDK Workshops*, pages 103–112.

Eckle-Kohler, J., Gurevych, I., Hartmann, S., Matuschek, M., and Meyer, C. M. (2012). Uby-lmf - a uniform model for standardizing heterogeneous lexical-semantic resources in iso-lmf. In *In Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC'12*, pages 275–282.

Eckle-Kohler, J., McCrae, J. P., and Chiarcos, C. (2015). Lemonuby–a large, interlinked, syntactically-rich lexical resource for ontologies. *Semantic Web*, 6(4):371–378.

Ehrmann, M., Cecconi, F., Vannella, D., Mccrae, J. P., Cimiano, P., and Navigli, R. (2014). Representing multilingual data as linked data: the case of babelnet 2.0. In *LREC*, pages 401–408.

Ehrmann, M., Jacquet, G., and Steinberger, R. (2016). Jrc-names: Multilingual entity name variants and titles as linked data. *Semantic Web*, 8:283–295.

Fang, Z., Wang, H., Gracia, J., Bosque-Gil, J., and Ruan, T. (2016). Zhishi. lemon: On publishing zhishi. me as linguistic linked open data. In *International Semantic Web Conference*, pages 47–55. Springer.

Ferrara, A. and Nikolov, A. (2011). Data linking for the semantic web. *Int. J. Semantic Web Inf. Syst.*, 7:46–76, 01.

Gonçalo Oliveira, H., Aguiar, F. S. d. S., and Rademaker, A. (2021). On the utility of word embeddings for enriching openwordnet-pt. In *3rd Conference on Language, Data and Knowledge (LDK 2021)*. Schloss Dagstuhl-Leibniz-Zentrum für Informatik.

Gracia, J., Montiel-Ponsoda, E., Cimiano, P., Gómez-Pérez, A., Buitelaar, P., and McCrae, J. (2012a). Challenges for the multilingual web of data. *Journal of Web Semantics*, 11:63–71.

Gracia, J., Montiel-Ponsoda, E., and Gomez-Perez, A. (2012b). Cross-lingual linking on the multilingual web of data (position statement). *CEUR Workshop Proceedings*, 936, 11.

Gracia, J., Villegas, M., Gomez-Perez, A., and Bel, N. (2018). The apertium bilingual dictionaries on the web of data. *Semantic Web*, 9(2):231–240.

Gracia, J., Kabashi, B., Kernerman, I., Lanau-Coronas, M., and Lonke, D. (2019). Results of the Translation Inference Across Dictionaries 2019 Shared Task. In Jorge Gracia, et al., editors, *Proceedings of TIAD-2019 Shared Task – Translation Inference Across Dictionaries, at 2nd Language, Data and Knowledge Conference (LDK 2019)*, volume 2493, pages 1–12, Sintra (Portugal). CEUR-WS.

Gracia, J., Fäth, C., Hartung, M., Ionov, M., Bosque-Gil, J., Veríssimo, S., Chiarcos, C., and Orlikowski, M. (2020). Leveraging linguistic linked data for cross-lingual model transfer in the pharmaceutical domain. In *International Semantic Web Conference*, pages 499–514. Springer.

Harris, S., Seaborne, A., and Prud'hommeaux, E. (2013). Sparql 1.1 query language. w3c recommendation (2013). *URL https://www. w3. org/TR/sparql11-query*.

Hartung, M., Orlikowski, M., and Verıssimo, S. (2020). Evaluating the impact of bilingual lexical resources on cross-lingual sentiment projection in the pharmaceutical domain.

Kapanadze, O. (2014). The multilingual grug parallel treebank–syntactic annotation for under-resourced languages. In *CCURL 2014: Collaboration and Computing for Under-Resourced Languages in the Linked Open Data Era*, Reykjavik, Iceland.

Kasner, Z. and Dušek, O. (2020). Train hard, finetune easy: Multilingual denoising for rdf-to-text generation. In *Proceedings of the 3rd International Workshop on Natural Language Generation from the Semantic Web (WebNLG+).*, pages 171–176.

Kernerman, I., Krek, S., Mccrae, J. P., Gracia, J., Ahmadi, S., and Kabashi, B. (2020). Introduction to the Globalex 2020 Workshop on Linked Lexicography. In Ilan Kernerman, et al., editors, *Proceedings of Globalex'20 Workshop on Linked Lexicography at LREC 2020*. ELRA.

Khan, A. F., Chiarcos, C., Declerck, T., Gifu, D., García, E. G.-B., Gracia, J., Ionov, M., Labropoulou, P., Mambrini, F., McCrae, J. P., et al. (accepted). When linguistics meets web technologies. recent advances in modelling linguistic linked open data. *Semantic Web Journal*.

Lanau-Coronas, M. and Gracia, J. (2020). Graph exploration and cross-lingual word embeddings for translation inference across dictionaries. In *Proceedings of the 2020 Globalex Workshop on Linked Lexicography*, pages 106–110, Marseille, France, May. European Language Resources Association.

Lehmann, J., Isele, R., Jakob, M., Jentzsch, A., Kontokostas, D., Mendes, P. N., Hellmann, S., Morsey, M., van Kleef, P., Auer, S., and Bizer, C. (2012). Dbpedia – a large-scale, multilingual knowledge base extracted from wikipedia. *Semantic Web*, 6(2):167–195.

Lesnikova, T. (2013). NLP for interlinking multilingual LOD. In *Proc. ISWC Doctoral consortium*, pages 32–39. No commercial editor.

Liu, Y., Gu, J., Goyal, N., Li, X., Edunov, S., Ghazvininejad, M., Lewis, M., and Zettlemoyer, L. (2020). Multilingual denoising pre-training for neural machine translation. *Transactions of the ACL*, 8:726–742.

Miller, G. A. (1995). WordNet: a lexical database for English. *Communications of the ACM*, 38(11):39–41.

Mititelu, V. B. and Mitrofan, M. (2020). The romanian medical treebank-simonero. In *Proceedings of the 15th International Conference on Linguistic Resources and Natural Language Processing Tools*, page 7.

Mitrofan, M. (2017). Bootstrapping a romanian corpus for medical named entity recognition. In *RANLP 2017 - Recent Advances in Natural Language Processing Meet Deep Learning*. Incoma Ltd. Shoumen, Bulgaria, nov.

Moran, S. and Chiarcos, C. (2020). 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*, page 39. MIT Press.

Moussallem, D., Sherif, M. A., Esteves, D., Zampieri, M., and Ngomo, A. N. (2018). LIDIOMS: A multilingual linked idioms data set. *CoRR*, abs/1802.08148.

Navigli, R. and Ponzetto, S. P. (2012a). BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial Intelligence*, 193:217–250.

Navigli, R. and Ponzetto, S. P. (2012b). Babelnet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial Intelligence*, 193:217–250.

Navigli, R. and Ponzetto, S. P. (2012c). Joining forces pays off: Multilingual joint word sense disambiguation. In *Proceedings of the 2012 joint conference on empirical methods in natural language processing and computational natural language learning*, pages 1399–1410. ACL.

Nentwig, M., Hartung, M., Ngomo, A. N., and Rahm, E. (2017). A survey of current Link Discovery frameworks. *Semantic Web*, 8(3):419–436.

Ngonga Ngomo, A.-C. and Auer, S. (2011). Limes - a time-efficient approach for large-scale link discovery on the web of data. In *Proceedings of IJCAI*.

Ngonga Ngomo, A.-C. (2012). On link discovery using a hybrid approach. *Journal on Data Semantics*, 1:203 – 217, 12.

Oliveira, H. G. (2021). Acquiring lexico-semantic knowledge from a portuguese masked language model. *Deep Learning and Neural Approaches for Linguistic Data*, page 13.

Pinnis, M. (2012). Latvian and lithuanian named entity recognition with tildener. *Seed*, 40:37.

Poornima, S. and Good, J. (2010). Modeling and encoding traditional wordlists for machine applications. In *Proceedings of the 2010 Workshop on NLP and Linguistics: Finding the Common Ground*, pages 1–9.

Rokas, A., Rackevičienė, S., and Utka, A. (2020). Automatic extraction of lithuanian cybersecurity terms using deep learning approaches. In *Human Language Technologies–The Baltic Perspective*, volume 328, pages 39–46. IOS Press.

Ruder, S., Vulić, I., and Søgaard, A. (2019). A survey of cross-lingual word embedding models. *Journal of Artificial Intelligence Research*, 65:569–631.

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

Saurí, R. and Grosse, J. (2021). Combining automatic and manual sense linking of dictionaries. `https://pret-a-llod.github.io/blog/combining-automatic-and-manual-sense-linking-of-dictionaries/`.

Sérasset, G. (2015). Dbnary: Wiktionary as a lemon-based multilingual lexical resource in RDF. *Semantic Web*, 6(4):355–361.

Speer, R., Chin, J., and Havasi, C. (2017). ConceptNet 5.5: An Open Multilingual Graph of General Knowledge. In *Proc AAAI Conference on Artificial Intelligence*, pages 4444–4451. `http://aaai.org/ocs/index.php/AAAI/AAAI17/paper/view/14972`.

Suchanek, F. M., Kasneci, G., and Weikum, G. (2007). Yago: a core of semantic knowledge. In *Proceedings of the 16th international conference on World Wide Web*, pages 697–706. ACM.

Swier, R. S. and Stevenson, S. (2005). Exploiting a verb lexicon in automatic semantic role labelling. In *Proceedings of the conference on HLT and EMNLP*, pages 883–890. ACL.

Tanaka, K. and Umemura, K. (1994). Construction of a Bilingual Dictionary Intermediated by a Third Language. In *Proc. COLING 1994*, pages 297–303.

Torregrosa, D., Arcan, M., Ahmadi, S., and McCrae, J. P. (2019). Tiad 2019 shared task: Leveraging knowledge graphs with neural machine translation for automatic multilingual dictionary generation. *Translation Inference Across Dictionaries*.

Villegas, M., Melero, M., Gracia, J., and Bel, N. (2016). Leveraging RDF Graphs for Crossing Multiple Bilingual Dictionaries. In Nicoletta Calzolari Conference Chair, et al., editors, *Proc. LREC 2016, Portorož (Slovenia)*, pages 868–876, Paris, France, may. ELRA.

Volz, J., Bizer, C., Gaedke, M., and Kobilarov, G. (2009). Silk–a link discovery framework for the web of data. *Proceedings of the 2nd Linked Data on the Web Workshop*, 01.

Wachowiak, L., Lang, C., Heinisch, B., and Gromann, D. (2021). Towards learning terminological concept systems from multilingual natural language text. In *3rd Conference on Language, Data and Knowledge (LDK 2021)*. Schloss Dagstuhl-Leibniz-Zentrum für Informatik.

Westphal, P., Stadler, C., and Pool, J. (2015). Countering language attrition with panlex and the web of data. *Semantic Web*, 6(4):347–353.

Wetzel, M. (2021). The journey to a multilingual sparql endpoint. `https://blog.coreon.com/2021/05/18/the-journey-to-a-multilingual-sparql-endpoint/`.

Witte, R. and Sateli, B. (2018). The lodexporter: Flexible generation of linked open data triples from nlp frameworks for automatic knowledge base construction. In *Proceedings of LREC 2018*, pages 2423–2428.

Xue, N. and Palmer, M. (2004). Calibrating features for semantic role labeling. In *Proceedings of EMNP 2004*.

8. Language Resource References

Gilles Sérasset. (2012–2022). *DBnary: 22 Wiktionary Language Editions as Lexical Linked Open Data*. Univ. Grenoble Alpes, ISLRN 023-163-901-149-4.