QuoteKG: A Multilingual Knowledge Graph of Quotes

Tin Kuculo[1(✉)], Simon Gottschalk[1], and Elena Demidova[2]

[1] L3S Research Center, Leibniz Universität Hannover, Hannover, Germany
{kuculo,gottschalk}@L3S.de
[2] Data Science and Intelligent Systems (DSIS), Universität Bonn, Bonn, Germany
demidova@cs.uni-bonn.de

Abstract. Quotes of public figures can mark turning points in history. A quote can explain its originator's actions, foreshadowing political or personal decisions and revealing character traits. Impactful quotes cross language barriers and influence the general population's reaction to specific stances, always facing the risk of being misattributed or taken out of context. The provision of a cross-lingual knowledge graph of quotes that establishes the authenticity of quotes and their contexts is of great importance to allow the exploration of the lives of important people as well as topics from the perspective of what was actually said. In this paper, we present QuoteKG, the first multilingual knowledge graph of quotes. We propose the QuoteKG creation pipeline that extracts quotes from Wikiquote, a free and collaboratively created collection of quotes in many languages, and aligns different mentions of the same quote. QuoteKG includes nearly one million quotes in 55 languages, said by more than 69,000 people of public interest across a wide range of topics. QuoteKG is publicly available and can be accessed via a SPARQL endpoint.

Keywords: Knowledge Graph · Quotes · Cross-lingual Alignment

1 Introduction

Quotes of public figures provide valuable information to understand their thoughts and attitudes, potentially leading to historically important actions, and thus serve as a crucial component in exploring world history [19]. Table 1 provides three examples of quotes, with the first one emphasising the relevance of historic quotes: in 1930, Winston Churchill recognised the value of reading them. The second example in Table 1 illustrates the relevance of quotes in world history: During a press conference in 2015, the German chancellor Angela Merkel

The original version of this chapter was revised: this chapter was previously published non-open access. The correction to this chapter is available at
https://doi.org/10.1007/978-3-031-06981-9_29

said "Wir schaffen das" ("We can do this") when the European migrant crisis unfolded and Germany prepared for the reception of refugees from Northern Africa and the Middle East. Since then, these three words defined Merkel's political course in the migrant crisis – and led both to a welcoming culture as well as the rise of nationalist protests and right-wing political parties [21,23].

Given this potential impact of words, it is of utmost importance to provide sources to quotes and to dismiss hoaxes [18,28]: The third example in Table 1 is a famous quote that has been attributed to different people, including Albert Einstein, Benjamin Franklin, and Mark Twain, but has not actually been said by any of them.[1] In general, a quote can be mentioned in different sources, and mentions can deviate. For example, "Wir schaffen das" might be mentioned as "We can do this" or "We will make it!" in English translations. Therefore, there is a need to align mentions to the same quote and to provide context information such as the source and description (e.g., "during a press conference").

Table 1. Three example quotes, together with their originators and dates. The last column gives examples of context that can be attributed to the mention of a quote, including source information, translations or validation of the quote's correctness.

Quote	By	Date	Selected context
It is a good thing for an uneducated man to read books of quotations. (en)	Winston Churchill	1930	*Source:* Roving Commission: My Early Life (1930) Chapter 9
Wir schaffen das. (de)	Angela Merkel	2015, Aug 31	*Translation:* We can do this (en)
The definition of insanity is doing the same thing over and over and expecting different results. (en)	Albert Einstein		Misattributed

In this paper, we introduce *QuoteKG* – a new knowledge graph that provides nearly one million quotes said by more than 69,000 persons of public interest in 55 languages. Quotes in *QuoteKG* come with detected sentiment and context such as their origin dates and sources. They are interlinked with their originators and other entities such as persons or events they refer to. Different mentions of the same quote are aligned across languages.

The creation of a knowledge graph covering quotes in many languages and their contexts faces several challenges detailed in the following.

– *Lack of context:* Most quote collections [10,24,34] lack context information and solely provide the quotes and their originators. To provide more context

---

[1] Reasons for false attribution of quotes to persons include to appear educated or to lend authority from the person. [27].

information in *QuoteKG*, we extract quotes from Wikiquote – a "free online compendium of sourced quotes from notable people"[2].

– *Tedious extraction process:* Even though Wikiquote is a semi-structured resource, extraction of quotes and contexts is a tedious process. In particular, we must design an extraction pipeline that is flexible across languages and adopts their characteristics. For example, it is necessary to differentiate the quotes not said by a person, but said about a person (e.g., English: "Quotes about Albert Einstein", German: "Zitate mit Bezug auf Albert Einstein").

– *Missing alignment of quote mentions:* As quote mentions in Wikiquote are not linked across languages, another important step is cross-lingual quote alignment which we perform using a language-agnostic transformer model that we evaluate on a ground truth set of manually aligned quote clusters.

Our contributions are as follows: (i) We propose a schema to represent quotes and context information. (ii) We propose an extraction pipeline that extracts quotes, their mentions and context information from all Wikiquote language versions. (iii) We align quote mentions across languages using a cross-lingual language model. (iv) We make *QuoteKG* publicly available[3].

The remainder of this paper is structured as follows: First, we describe the impact of *QuoteKG* in the fields of Semantic Web, Natural Language Processing, Digital Humanities and others in more detail. Then, in Sect. 3, we describe the schema adopted for *QuoteKG*. In Sect. 4, we describe the *QuoteKG* creation pipeline. In Sect. 5, we provide statistics and examples of *QuoteKG*, followed by information about the availability and maintenance in Sect. 6. Section 7 gives an overview of related work. Finally, we provide a conclusion in Sect. 8.

2   Potential Impact

QuoteKG contains quotes, a new type of information that is, to the best of our knowledge, not yet present in existing knowledge graphs. Therefore, *Quote-KG* can potentially attract new audiences from several fields such as Digital Humanities and Natural Language Processing. While existing cross-domain or event-centric knowledge graphs such as Wikidata [36], DBpedia [2], and EventKG [12] target the representation of real-world entities, including persons of public interest and important events, they scarcely represent what people actually said – even though this information reflects persons' characteristics and can lead to an understanding of how particular events unfolded in the real world. Instead, facts about persons in knowledge graphs (e.g., properties representing birth dates, marriages, and awards received) without a doubt represent relevant facts in a person's life but typically do not reveal personal traits or surprising insights. Existing corpora of quotes like Quotebank [34] and the QUOTES500K dataset [10] provide large collections of English quotes. In contrast to these corpora, *QuoteKG* is a knowledge graph and provides societal relevant quotes and

---

[2] https://en.wikiquote.org/wiki/Main_Page.

[3] https://www.quotekg.l3s.uni-hannover.de.

contexts in 55 languages, links them to other knowledge graphs, and aligns quote mentions across languages.

Potential applications of *QuoteKG* are manifold: (i) First and foremost, *QuoteKG* can add a new dimension to the exploration and investigation of the lives of public figures. While the creation of biography timelines from knowledge graphs has been studied in the past [1,13], such timelines do not consider the inclusion of quotes. *QuoteKG* can help to enrich such timelines with relevant quotes to make them more lively and informative. (ii) Similarly, the analysis of quotes related to a specific topic over time can support research in the fields of Digital Humanities, and can be used to gauge public opinion regarding specific events. For example, there have been analyses of how social movements and global events affect language [14] and how the words used by public persons carry political backgrounds [35]. (iii) Quotes also play an important role when observing information propagation [31] and the bias potentially caused by one-sided selection of quotes [25]. (iv) *QuoteKG* can also serve as an additional resource for machine translation, given that it contains 38,931 quotes with mentions in different languages. (v) *QuoteKG* can help answering questions such as "Who said 'yes, we can'?". (vi) *QuoteKG* contains 13,104 quotes labelled as misattributed or falsely claimed. They can be used as a resource for understanding the propagation of false or misleading information [28]. (vii) Finally, *QuoteKG* can take quote collections to a new level: There are plenty of websites available that provide collections of quotes[4], typically monolingual and primarily for entertainment purposes (e.g., images of inspirational quotes or quote mashup games[5]). The context information in *QuoteKG* can support the exploration and the search for quotes and provide important information surrounding a quote, and, as such, broaden the user's horizon.

3   *QuoteKG* Schema

The goal of the *QuoteKG* schema is to model quotes, their relationships with persons and other entities, as well as their different mentions, e.g. translations, typically in different contexts. To this end, *QuoteKG* is based on an extension of the *schema.org* vocabulary that provides a `so:Quotation`[6] class which is reused. According to the schema.org description, the `so:Quotation` class models quotes that are "Often but not necessarily from some written work" and can also refer to a "Quotation from an Event"[7]. Therefore, it fits well to our concept of a quote in *QuoteKG*. However, we extend the schema with a new class `qkg:Mention` which models the different mentions of a quote.
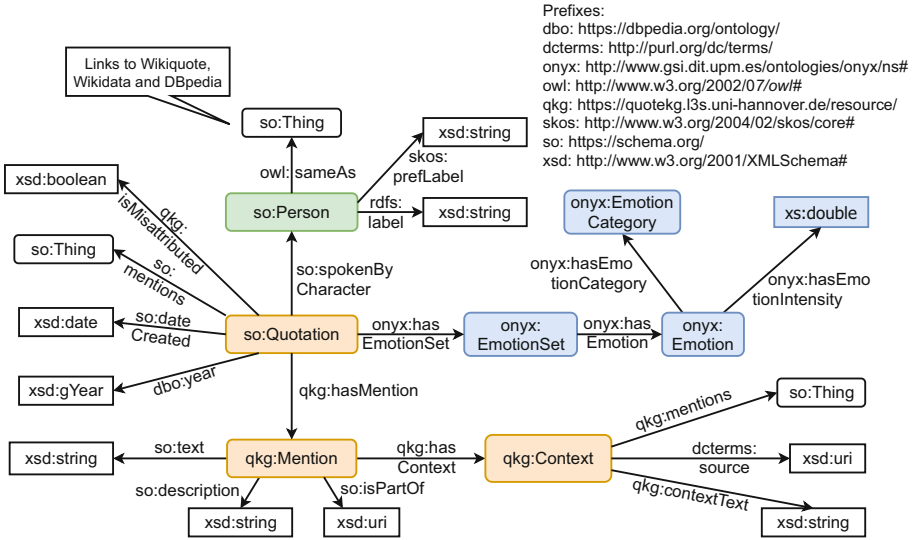
Figure 1 presents *QuoteKG*'s schema. Its classes are described in the following.

---

[4] https://www.brainyquote.com, https://www.goodreads.com/quotes, https://www.successories.com/iquote, ...

[5] http://natetyler.github.io/.

[6] so: https://schema.org/.

[7] https://schema.org/Quotation.

Fig. 1. The *QuoteKG* schema based on *schema.org*. Arrows visualize the `rdfs:domain` and `rdfs:range` restrictions on properties. Name spaces and prefixes are described in the top right corner. Orange classes are related to quotes and their mentions, blue ones to the sentiment of a quote and the green class is about the person. (Color figure online)

- **Person:** Each quote in *QuoteKG* is assigned to a person modeled as `so:Person`. For persons, *QuoteKG* provides additional type information (e.g., Politician) plus `owl:sameAs` relations to Wikidata and the different DBpedia and Wikiquote language editions.
- **Quote:** In *QuoteKG*, a resource typed as `so:Quotation` refers to the unique event of something being said by a person of public interest (`so:spokenByCharacter`) at a specific point in time (`so:dateCreated`). A quote may also refer to other entities (`so:mentions`) of any type.
- **Mention:** A quote can be mentioned in different contexts: For example, there may be translations of the quote in different languages, alternative records of the same quote, or different contexts that a quote is extracted from. Therefore, we introduce the class `qkg:Mention`. Mentions can be related to one or more `qkg:Context` objects.
- **Context:** The context of a mention provides additional attributes that come together with the specific mention. For example, its origin (e.g., a reference to a specific interview) and the original source (e.g., a link to a news website). To model context, we create the class `qkg:Context`.
- **Sentiment:** For each quote, we provide its sentiment using the Onyx ontology which is used for describing emotions [29]. A quote is assigned a score for a specific emotion category ("neutral", "negative" or "positive").
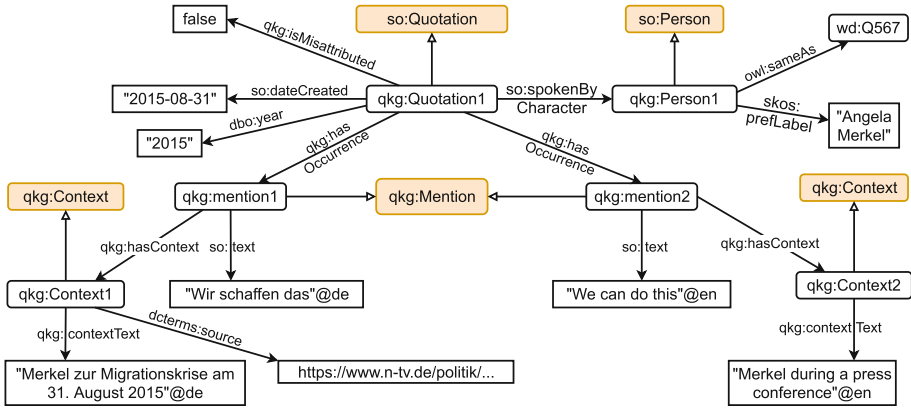
Fig. 2. An example quote modeled using the *QuoteKG* schema. ⇢ marks `rdf:type` relations. `xsd:data` type annotations were omitted for brevity. The prefixes and name spaces are the same as in Fig. 1, plus `wd:` https://www.wikidata.org/entity/.

Figure 2 shows an example instantiation of the *QuoteKG* schema. The quote "Wir schaffen das" introduced in Table 1 is connected to two instances of `qkg:Mention`, one representing a German mention, the other one an English one ("We can do this"). Both mentions come with additional context information.

4   Extraction and Alignment of Quotes

This section describes the input data and the implementation of the four main steps of the *QuoteKG* creation pipeline shown in Fig. 3.
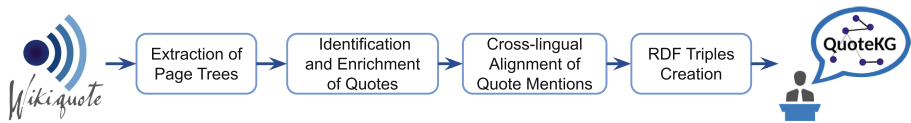
Fig. 3. Pipeline to create *QuoteKG* from Wikiquote.

4.1   Wikiquote

We base *QuoteKG* on Wikiquote – an online collection of quotes[8]. Wikiquote has a similar structure to Wikipedia: Independent versions of Wikiquote exist for different languages. Wikiquote contains pages, each of them about a given topic and divided into different sections and subsections. For *QuoteKG*, we focus on

---

[8] https://en.wikiquote.org/wiki/Main_Page.

```
*  '''Falling in love is not at all the most stupid thing that people do
|  but gravitation cannot be held responsible for it.'''
** Jotted (in German) on the margins of a letter to him (1933), p. 56
```

Fig. 4. Example of a quote in the English Wikiquote, based on MediaWiki markup.

```
{{Citation|Tomber amoureux n'est pas du tout la chose la plus stupide
que font les gens | mais la gravitation ne peut en être tenue pour
responsable. |original=Falling in love is not at all the most stupid
thing that people do | but gravitation cannot be held responsible for
it.|langue=en}}
```

Fig. 5. Example of a quote in the French Wikiquote, using templates.

Wikiquote pages about persons that contain quotes attributed to them. Example pages are the English[9] and French page about Albert Einstein[10].

Each Wikiquote page is formatted using the MediaWiki markup[11] and contains semi-structured content that includes the person's description, sections with quotes, references and more. The quotes are given in one of the following representations: in the traditional MediaWiki markup as shown in Fig. 4 or using pre-defined *templates* that allow for a more structured definition of key-value pairs. For example, Fig. 5 shows the key-value pair (*key*: `Citation`, *value*: `Tomber amoureux...`).

While there are links between the pages describing persons in different languages, quote mentions are not linked across languages. Figure 4 and Fig. 5 show two mentions of the same quote by Albert Einstein. The first is from the English Wikiquote and shows an English quote, the second one from the French Wikiquote is given in French and English. The original German quote is not available in these two language versions.

In general, one can observe a large imbalance in Wikiquote regarding the covered persons and the number of quotes in different language versions. This imbalance can often be explained by the different sizes of Wikiquote language versions and the difference in the cultural significance of a person in one language community compared to another. For example, there exists a French page with 35 quotes and an Italian page with 2 quotes of the former footballer Michel Platini who used to play in Italy, but there is no English page. This imbalance also implies that there is no guarantee that Wikiquote will contain the original language version of a quote. *QuoteKG* can have multiple quote mentions of the same quote through cross-lingual quote mention alignment.

---

[9] https://en.wikiquote.org/wiki/Albert_Einstein.
[10] https://fr.wikiquote.org/wiki/Albert_Einstein.
[11] https://www.mediawiki.org/wiki/Help:Formatting.
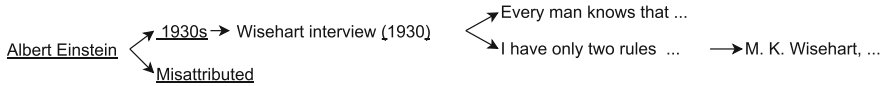
Fig. 6. Excerpt of an example page tree from the English Wikiquote page about Albert Einstein. Section titles are underlined.

4.2 Extraction of Page Trees

In the beginning, our *QuoteKG* creation pipeline processes all Wikiquote language editions with at least 50 pages, excluding Simple English[12] and selects all pages about persons. From each Wikiquote page about a person, we create a *page tree*. The page tree consists of section titles plus quotes and contexts. An example page tree is presented in Fig. 6.

4.3 Identification and Enrichment of Quotes

In the second step of the *QuoteKG* creation pipeline, the page trees are transformed into a set of quotes with contextual information. To this end, we specify language-specific rules and enrich quotes and contexts with additional metadata.

To identify quotes, we first define a language-specific list of section titles denoting quotes (e.g., "Citations", "Zitate", "Citazioni") and contextual information (e.g., "útskýring", "Viitattu", "vydavatel"). In addition, we collect a list of template types representing quotes and consider all child nodes of section titles as quotes. From section titles and templates, we further gather the following:

– Dates: We identify the dates of quotes from a pre-defined list of template keys (e.g., "année d'origine" in French) for quotes extracted from templates. If such dates are not available or when dealing with quotes not extracted from templates, we extract dates from the section titles above the particular quote in the page tree and the contexts below the quote.
 We select the time expression with the highest level of precision (e.g., we select May 2020 over 2020). In case of conflicts, no date is chosen.
– Veracity: To reflect the authenticity of quotes and their contextual information, we capture whether a quote has been misattributed to the person. In Wikiquote, misattributed quotes are grouped into specified sections. We identify such sections with a manually created list of regular expressions (e.g., "Misattributed" (English) and "Fälschlich zugeschrieben" (German).
– Sources: Often, context contains links to websites where the quote was reported. We collect such external links from templates and from the markup.
– Linked entities: Quotes can be linked to entities such as other persons or organisations. We collect such links from templates and the markup.

---

[12] For more detailed statistics about Wikiquote language editions see: https://wikistats.wmcloud.org/display.php?t=wq.

- Language: While the Wikiquote pages are written in specific languages, their quotes can be written in their original language or translated. For this reason, we use language detection to designate the language of a quote and do not rely on the language of the page tree.
- Sentiment: We detect the sentiment of each quote mention (*positive*, *negative* or *neutral* with a score between 0 and 1) using XLM-RoBERTa-Twitter, an XLM-RoBERTa model trained on ∼198 M multilingual tweets [3].
- Identity links: To establish `owl:sameAs` links between the QuoteKG entities, Wikidata and DBPedia, we use Wikidata's sitelinks[13].

For all persons and entities identified during this process, we extract additional information regarding their labels and types from DBpedia and Wikidata.

4.4   Cross-Lingual Alignment of Quote Mentions

After identifying and enriching quotes, we need to detect which of them represent mentions of the same quote said by a person of public interest. This task of cross-lingual alignment of quote mentions is treated as a clustering task at the end of which each cluster represents a quote with a set of mentions.

In detail, the clustering task is performed for each person in isolation. Given a person's quote mentions in a set of languages, we aim at creating clusters of highly similar mentions. To derive a similarity between two mentions, potentially from different languages, we compute the cosine similarity of sentence embeddings derived from the mentions' texts. As an embedding model, we use a language-agnostic transformer model pre-trained on millions of multilingual paraphrase examples in more than 30 languages, namely XLM-RoBERTa [8]. The ability of such models to adapt to previously unknown languages has been shown in [16]. Given such embeddings and the cosine similarity function, clustering is performed by detecting communities of quotes using a nearest-neighbour search. To do so, we chose UKPLab's Fast Clustering algorithm[14] that is optimised towards efficient similarity computations of our embeddings.

To aggregate the sentiments of all mentions in a cluster, we take the most frequent sentiment category and average over the scores of that category.

4.5   RDF Triples Creation

After identification of quotes and their contexts and cross-lingual alignment, we transform them into RDF triples following the schema presented in Fig. 1.

4.6   Implementation

We use the MWDumper[15] to process the Wikiquote XML dumps and parse the single pages given in the Wikipedia markup using the Bliki engine[16]. For lan-

---

[13] https://www.wikidata.org/wiki/Help:Sitelinks/en-gb.
[14] https://github.com/UKPLab/sentence-transformers/blob/master/examples/applications/clustering/.
[15] https://www.mediawiki.org/wiki/Manual:MWDumper.
[16] https://github.com/axkr/info.bliki.wikipedia_parser.

Table 2. Statistics of selected languages in *QuoteKG*.

Language	Persons	Quotes	Mentions	Mentions with contexts
English	19,073	267,740	271,541	193,848
Italian	18,803	145,235	146,103	48,107
German	3,461	16,012	16,441	4,330
Croatian	2,707	11,023	12,965	2,045
Welsh	239	461	508	247
All Languages	69,467	880,878	961,535	411,912

guage detection and time expression extraction, we use the langdetect[17] and dateparser[18] libraries. The Fast Clustering algorithm was run with a cosine similarity threshold of 0.8. The creation of knowledge graph triples and their serialisation is done via the RDFLib library[19]. The Java implementation of the dumper and the Python code for cross-lingual alignment and knowledge graph creation are publicly available on GitHub[20].

5   Statistics, Evaluation, Examples and Web Interface

In this section, we first provide general statistics of *QuoteKG*, evaluate the cross-lingual alignment and present example queries.

5.1   Statistics

In total, *QuoteKG* contains $880,878$ quotes with $961,535$ quote mentions. For $411,912$ mentions, context is available. Table 2 provides detailed statistics for selected languages. *QuoteKG* covers both high-resource languages such as English ($271,541$ quote mentions from $19,073$ persons) and Italian ($146,103$ quote mentions from $18,803$ persons), as well as low-resource languages such as Welsh (508 quote mentions from 239 persons).

5.2   Evaluation of the Cross-Lingual Alignment

We evaluate the quality of the cross-lingual alignment of quote mentions by comparing to a ground truth of correctly clustered mentions. Creating such a ground truth is a tedious process due to the large amount of possible clusterings

---

[17] https://pypi.org/project/langdetect/.

[18] https://github.com/scrapinghub/dateparser/.

[19] https://github.com/RDFLib/rdflib.

[20] https://github.com/tkuculo/QuoteKG.

and the number of pairwise comparisons[21]. We have selected eight persons with quotes in English, German and Italian and manually clustered their mentions. Ground truth clusters were then compared to the *QuoteKG* clusters by viewing the clustering process as a series of decisions for each of the pairs of mentions [30]. For example, we consider three positive pairs for a quote mentioned in three languages: (Mention$_1$, Mention$_2$), (Mention$_1$, Mention$_3$), (Mention$_2$, Mention$_3$).

Table 3 shows the results of this evaluation: Cross-language alignment in *QuoteKG* shows an average precision of 1.0 and an F$_1$ score of 0.99 for this ground truth data set. Following the imbalance of Wikiquote's coverage described in Sect. 4.1, there is a high number of true negatives, i.e., the majority of quotes are only mentioned once in all Wikiquote language versions. In total, there are only two mentions which are not clustered together but should have been. All the other clusters are correct.

Table 3. Evaluation of cross-lingual alignment for eight selected persons in English, German and Italian. TP: true positives (mention pairs that were correctly clustered together) TN: true negatives (mention pairs that were correctly not clustered together), FP: false positives, FN: false negatives, P: precision, R: recall, F$_1$: F$_1$ score.

Person	TP	TN	FP	FN	P	R	F$_1$
Alan Turing	10	935	0	1	1	0.91	0.95
Alexander the Great	5	491	0	0	1	1.0	1.0
Edward Snowden	6	697	0	0	1	1.0	1.0
Gustav Mahler	1	44	0	0	1	1.0	1.0
Jean-Claude Juncker	4	776	0	0	1	1.0	1.0
Marie Antoinette	4	347	0	0	1	1.0	1.0
Marie Curie	2	251	0	0	1	1.0	1.0
Tom Clancy	1	2849	0	0	1	1.0	1.0
Total	33	6390	0	1	1.0	0.99	0.99

Our ground truth set of manually aligned quote clusters is available on the *QuoteKG* website.

5.3 Example Queries

In this section, we present two example queries demonstrating how to use *Quote-KG* as a collection of quotes and as a resource to conduct research on the misattribution of quotes.

---

[21] When considering a person that has 10 quotes in 5 languages each, there are $\sum_i^{5-1} 10 \cdot i^2 = 1,000$ possible pairwise comparisons.

```
SELECT ?Person (COUNT(?quote) AS ?NumberOfQuotes) WHERE {
  ?quote a so:Quotation ;
    so:spokenByCharacter [
      skos:prefLabel ?Person ] .
} GROUP BY ?Person
ORDER BY DESC(COUNT(?quote))
```

Listing 1.1. SPARQL query: Persons with the most quotes.

***QuoteKG* as a Collection of Quotes and their Originators.** Listing 1.1 shows a SPARQL query that returns the five persons with the most quotes in *QuoteKG*. Table 4 shows these persons together with the number of quotes. Without surprise, the persons with most quotes are philosophers and writers, including Friedrich Nietzsche and Oscar Wilde, plus Albert Einstein, known for many (misattributed) quotes [28].

Table 4. The first five results of the query in Listing 1.1.

?Person	?NumberOfQuotes
Friedrich Nietzsche	2,530
Oscar Wilde	1,786
Albert Einstein	1,627
Donald Trump	1,610
Johann Wolfgang von Goethe	1,537

Verification of Quotes. Misinformation on the Internet has become an increasingly important problem and requires methods that classify the veracity of information [33] and benefit from knowledge graphs such as ClaimsKG that provide annotated and erroneous facts [32]. While ClaimsKG provides wrong claims stated by persons extracted from fact-checking sites, *QuoteKG* has quotes labelled as wrongly attributed to persons, thus a different type of misinformation. The query shown in Listing 1.2 returns quotes of Albert Einstein that are marked as misattributed in *QuoteKG* (see Table 5), together with context information. Such context information can be a valuable resource for explaining misattribution in the case of quotes.

5.4   Web Interface

On the *QuoteKG* website, we offer a SPARQL endpoint and a demo Search & Demo interface where users can search for specific persons and display their quotes in selected languages. An example of this interface is shown in Fig. 7 which display Portuguese and English quotes of Johann Wolfgang von Goethe.

```
SELECT ?Text (SAMPLE(?contextText) AS ?contextTexts)
(SAMPLE(?source) AS ?Source) WHERE {
  ?quote so:spokenByCharacter [
    skos:prefLabel "Albert_Einstein" ] ;
  qkg:isMisattributed true ; qkg:hasMention ?mention .

  ?mention so:text ?Text ;
    qkg:hasContext [
    qkg:contextText ?contextText ; so:source ?source
  ] .
} GROUP BY ?Text
```

Listing 1.2. SPARQL query: Quotes misattributed to Albert Einstein and their contexts.

Table 5. Two results of the query in Listing 1.2, returning quotes that were misattributed to Albert Einstein. Texts are shortened for brevity here.

?Text	?contextTexts	?Source
Everything is energy and that's all there is to it . . . It can be no other way. This is not philosophy. This is physics	There's no evidence that Einstein ever said this	http:// quoteinvestigator. com/2012/05/16/ everything-energy/
If the facts don't fit the theory, change the facts	The earliest published attribution of this quote to Einstein found on . . . , but no source to Einstein's original writings is given . . .	http://books.google. com/books?id=...

6   Availability

Availability: The *QuoteKG* website[22] provides access to a description of *QuoteKG* and its schema, to the SPARQL endpoint, to data downloads and will provide a canonical citation to this paper. *QuoteKG* is licensed under the Creative Commons Attribution Share Alike 4.0 International[23] license. Persistent access to the *QuoteKG* triple files is provided through an upload to the zenodo repository[24]. The code for the creation of *QuoteKG* is publicly available on GitHub[25] and is licensed under the MIT license[26].

Sustainability Plan: To account for updates in all Wikiquote language editions, we plan to release new versions of *QuoteKG* twice a year. To do so, we deploy a script that covers the entire pipeline depicted in Fig. 3, from the download of Wikiquote dumps in each language until the creation of triples.

---

[22] https://quotekg.l3s.uni-hannover.de.
[23] https://creativecommons.org/licenses/by-sa/4.0/legalcode.
[24] https://zenodo.org/record/4702544.
[25] https://github.com/tkuculo/QuoteKG.
[26] https://opensource.org/licenses/MIT.

| 1832 (RDF) | • EN: Once a man's thirty, he's already old, He is indeed as good as dead. It's best to kill him right away. (RDF) |
| 1832 (RDF) | • EN: Law is mighty, mightier necessity. (RDF)
 • PT: A lei é poderosa; mais poderosa, porém, é a necessidade. (RDF) |

Fig. 7. An example of the Search & Demo interface showing two quotes of Johann Wolfgang Goethe which are available in Portuguese or English. The sentiment of quotes is indicated by colour (red: negative, green: positive). (Color figure online)

Adherence to Standards: *QuoteKG* is modeled through the Resource Description Framework. Its schema is an extension of *schema.org*. We provide a machine-readable description of *QuoteKG* using the VoID vocabulary[27]. *QuoteKG* adheres to the Linked Data Principles: resources can be looked up through their URIs and they are interlinked with Wikidata and DBpedia.

7   Related Work

In this section, we give an overview of other corpora and knowledge graphs containing quotes, other usages of Wikiquote and about cross-lingual alignment.

Quote Corpora. Many collections of quotes have been created and maintained, mainly mono-lingual and without semantic annotations. Since the release of its first edition in 1941, The Oxford Dictionary of Quotations [20] aims at providing "the wit and wisdom of past and present" with a focus on the provenance of quotes. Provenance of quotes is also an indispensable criterion in the Book of Fake Quotes [4]. There are few machine-readable monolingual quote collections[28,29] [10,24,34]. These corpora are typically monolingual and extracted from news. Consequently, while they may have a large amount of quotes, they lack a mechanism to ensure societal relevance of quotes as in Wikiquote. As a knowledge graph, *QuoteKG* enables easy access to quotes and rich metadata.

Quotes in Knowledge Graphs. While DBQuote [26] allows user annotations of quotes extracted from Twitter and Wikiquote through an ontology, it only covers two languages (English and Korean) and has not been made available. To the best of our knowledge, *QuoteKG* is the first publicly available knowledge graph of quotes. Consequently, quotes have only been insufficiently covered in the Semantic Web: for example, Wikidata [36] contains less than 400 instances of the class "Phrase"[30] that are attributed to an author or creator – most of them only consisting of few words (e.g., "cogito ergo sum" and "covfefe"). Event-centric knowledge graphs such as EventKG [12] provide an understanding of the human history and world-shaking events. They do not include quotes that complement

---

[27] https://www.w3.org/TR/void/.
[28] https://www.kaggle.com/akmittal/quotes-dataset.
[29] https://github.com/JamesFT/Database-Quotes-JSON.
[30] https://www.wikidata.org/wiki/Q187931.

the deeds of public figures. Many applications based on knowledge graphs (e.g., for exploring the lives of persons of public interest [1,13]) could immediately profit from the inclusion of quotes.

Wikiquote. Until now, Wikiquote has rarely been used as a research corpus, presumably due to the required but tedious extraction process. One example is the work by Buscaldi et al. who manually tagged quotes of the Italian Wikiquote as humorous or not, and used their annotated corpus for training models for humour recognition [5]. Giammona et al. analysed the spread of ancient quotes in today's Web through Wikiquote [9] and Wikiquote was used for training the chatbot Poetwannabe [6]. With *QuoteKG*, we foresee to ease the access to quotes for a wide range of research questions.

Cross-Lingual Alignment. Several studies have shown that different languages share similar statistical properties that can be used to learn cross-lingual alignments between two languages, even without relying on any form of bilingual supervision [7]. While most works and datasets address bilingual alignment [11,15,17], there are only few works on cross-lingual alignment [22]. *QuoteKG* focuses on the specific task of cross-lingual alignment of quote mentions.

8   Conclusion

In this paper, we presented *QuoteKG* – a novel, multilingual knowledge graph of quotes. We have presented the *QuoteKG* schema based on schema.org as well as a pipeline that extracts quotes from the Wikiquote corpus and aligns them across languages. *QuoteKG* is publicly available and includes nearly one million quotes quotes in 55 languages, said by nearly 69,000 people of public interest.

References

1. Althoff, T., Dong, X.L., Murphy, K., Alai, S., Dang, V., Zhang, W.: TimeMachine: timeline generation for knowledge-base entities. In: Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 19–28 (2015)
2. Auer, S., Bizer, C., Kobilarov, G., Lehmann, J., Cyganiak, R., Ives, Z.: DBpedia: a nucleus for a web of open data. In: Aberer, K., et al. (eds.) ASWC/ISWC -2007. LNCS, vol. 4825, pp. 722–735. Springer, Heidelberg (2007). https://doi.org/10.1007/978-3-540-76298-0_52
3. Barbieri, F., Anke, L.E., Camacho-Collados, J.: XLM-T: A Multilingual Language Model Toolkit for Twitter. arXiv preprint arXiv:2104.12250 (2021)
4. Boller, P.F., Jr., George, O.J., Jr., et al.: They Never Said It: A Book of Fake Quotes, Misquotes, and Misleading Attributions: A Book of Fake Quotes, Misquotes, and Misleading Attributions. Oxford University Press, Oxford (1989)

5. Buscaldi, D., Rosso, P.: Some experiments in humour recognition using the Italian Wikiquote collection. In: Masulli, F., Mitra, S., Pasi, G. (eds.) WILF 2007. LNCS (LNAI), vol. 4578, pp. 464–468. Springer, Heidelberg (2007). https://doi.org/10.1007/978-3-540-73400-0_58

6. Chorowski, J., Lancucki, A., Malik, S., Pawlikowski, M., Rychlikowski, P., Zykowski, P.: A talker ensemble: the University of Wroclaw's entry to the NIPS 2017 conversational intelligence challenge. In: Escalera, S., Weimer, M. (eds.) The NIPS '17 Competition: Building Intelligent Systems. TSSCML, pp. 59–77. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-94042-7_4

7. Chung, Y.A., Weng, W.H., Tong, S., Glass, J.: Unsupervised Cross-Modal Alignment of Speech and Text Embedding Spaces. arXiv preprint arXiv:1805.07467 (2018)

8. Conneau, A., et al.: Unsupervised Cross-Lingual Representation Learning at Scale. arXiv preprint arXiv:1911.02116 (2019)

9. Giammona, C., Yanes, E.S.: From Print to Digital Texts, from Digital Texts to Print. Indirect Tradition of Latin Classics on the Web. Storie e Linguaggi 5(1) (2019)

10. Goel, S., Madhok, R., Garg, S.: Proposing contextually relevant quotes for images. In: Pasi, G., Piwowarski, B., Azzopardi, L., Hanbury, A. (eds.) ECIR 2018. LNCS, vol. 10772, pp. 591–597. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-76941-7_49

11. Gottschalk, S., Demidova, E.: MultiWiki: interlingual text passage alignment in Wikipedia. ACM Trans. Web **11**(1), 6:1–6:30 (2017)

12. Gottschalk, S., Demidova, E.: EventKG-the hub of event knowledge on the web-and biographical timeline generation. Semant. Web **10**(6), 1039–1070 (2019)

13. Gottschalk, S., Demidova, E.: EventKG+BT: generation of interactive biography timelines from a knowledge graph. In: Harth, A., et al. (eds.) ESWC 2020. LNCS, vol. 12124, pp. 91–97. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-62327-2_16

14. Haun, M.: How social movements and global events are changing language in 2020 (2020). Accessed 03 Dec 2021

15. Hieber, F.: WikiCLIR: A Cross-lingual Retrieval Dataset from Wikipedia. Universität (2014)

16. Hu, J., Ruder, S., Siddhant, A., Neubig, G., Firat, O., Johnson, M.: XTREME: a massively multilingual multi-task benchmark for evaluating cross-lingual generalisation. In: International Conference on Machine Learning, pp. 4411–4421. PMLR (2020)

17. Jing, Y., Xiong, D., Zhen, Y.: BiPaR: A Bilingual Parallel Dataset for Multilingual and Cross-Lingual Reading Comprehension on Novels. arXiv preprint arXiv:1910.05040 (2019)

18. Keyes, R.: The Quote Verifier: Who Said What, Where, and When. St. Martin's Griffin (2007)

19. Khurana, S.: These 4 Quotes Completely Changed the History of the World. https://www.thoughtco.com/quotes-that-changed-history-of-world-2831970. Accessed 01 Dec 2021

20. Knowles, E.: The Oxford Dictionary of Quotations. Oxford University Press, Oxford (2009)

21. Krämer, A.: Ein Satz mit Folgen (2021). https://www.tagesschau.de/inland/merkel-wir-schaffen-das-109.html. Accessed 01 Dec 2021

22. Liang, Y., et al.: XGLUE: A New Benchmark Dataset for Cross-Lingual Pretraining, Understanding and Generation. arXiv preprint arXiv:2004.01401 (2020)

23. Mushaben, J.M.: Wir schaffen das! Angela Merkel and the European refugee crisis. German Polit. **26**(4), 516–533 (2017)
24. Newell, C., Cowlishaw, T., Man, D.: Quote extraction and analysis for news. In: Proceedings of the Workshop on Data Science, Journalism and Media, KDD, pp. 1–6 (2018)
25. Niculae, V., Suen, C., Zhang, J., Danescu-Niculescu-Mizil, C., Leskovec, J.: QUOTUS: the structure of political media coverage as revealed by quoting patterns. In: Proceedings of the 24th International Conference on World Wide Web, pp. 798–808 (2015)
26. Piao, G., Breslin, J.G.: DBQuote: a social web based system for collecting and sharing wisdom quotes. In: Proceedings of the 5th Joint International Semantic Technology Conference, Poster and Demonstrations (2015)
27. Reucher, G.: Famos Quotes: Why are so many fake? (2021). Accessed 03 Dec 2021
28. Robinson, A.: Did Einstein really say that? Nature **557**(7703), 30–31 (2018)
29. Sánchez-Rada, J.F., Iglesias, C.A.: Onyx: a linked data approach to emotion representation. Inf. Process. Manag. **52**(1), 99–114 (2016)
30. Schütze, H., Manning, C.D., Raghavan, P.: Introduction to Information Retrieval, vol. 39. Cambridge University Press, Cambridge (2008)
31. Sims, M., Bamman, D.: Measuring information propagation in literary social networks. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP) (2020)
32. Tchechmedjiev, A., et al.: ClaimsKG: a knowledge graph of fact-checked claims. In: Ghidini, C., et al. (eds.) ISWC 2019. LNCS, vol. 11779, pp. 309–324. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-30796-7_20
33. Thorne, J., Vlachos, A.: Automated fact checking: task formulations, methods and future directions. In: Proceedings of the 27th International Conference on Computational Linguistics, pp. 3346–3359 (2018)
34. Vaucher, T., Spitz, A., Catasta, M., West, R.: Quotebank: a corpus of quotations from a decade of news. In: Proceedings of the 14th ACM International Conference on Web Search and Data Mining, pp. 328–336 (2021)
35. Viala-Gaudefroy, J., Lindaman, D.: Donald Trump's 'Chinese Virus': The Politics of Naming (2020). Accessed 03 Dec 2021
36. Vrandečić, D., Krötzsch, M.: Wikidata: a free collaborative knowledge base. Commun. ACM **57**(10), 78–85 (2014)