Undefined 1 (2009) 1–5
IOS Press

Semantic Quran

A Multilingual Resource for Natural-Language Processing

Editor(s): Name Surname, University, Country
Solicited review(s): Name Surname, University, Country
Open review(s): Name Surname, University, Country

Mohamed Ahmed Sherif [a] and Axel-Cyrille Ngonga Ngomo [a]
[a] *Universität Leipzig, Institut für Informatik, AKSW, Postfach 100920, D-04009 Leipzig, Germany*
E-mail: {lastname}@informatik.uni-leipzig.de

Abstract. In this paper we describe the Semantic Quran dataset, a multilingual RDF representation of translations of the Quran. The dataset was created by integrating data from two different semi-structured sources. The dataset were aligned to an ontology designed to represent multilingual data from sources with a hierarchical structure. The resulting RDF data encompasses 43 different languages which belong to the most under represented languages in Linked Data, including Arabic, Amharic and Amazigh. We designed the dataset to be easily usable in natural-language processing applications with the goal of facilitating the development of knowledge extraction tools for these languages. In particular, the Semantic Quran is compatible with the Natural-Language Interchange Format and contains explicit morpho-syntactic information on the utilized terms. We present the ontology devised for structuring the data. In this paper, we also provide the transformation rules implemented in our extraction framework. Finally, we detail the link creation process as well as possible usage scenarios for the Semantic Quran dataset.

Keywords: Multilingual dataset, Natural Language Processing, Morpho-syntatic data, Linked Data, ontology

1. Introduction

Over the last years, the Linked Open Data (LOD) movement has gained significant momentum [1]. A large number of datasets was extracted from sources as different as Wikipedia infoboxes and curated biomedical databases. Still, most of the datasets in the Linked Data Cloud contain only English labels and fail to represent the diversity of languages used across the Web.[1] Yet, a more multilingual Linked Data Cloud would represent a tremendous resource that can be used for novel knowledge extraction techniques and more broadly for novel natural-language processing (NLP) approaches. For example, novel NLP approaches for minority languages could be developed by reusing information available across the different languages [6]. Moreover, a structured representation of corpora would improve their use in applications such as the specification of templates for question answering [7] or the efficient merging with other resources [3].

In this paper, we present the *Semantic Quran dataset*. We aim to contribute towards the realization of the vision of a multilingual Linked Data Cloud and thus support the adaptation of NLP tools for languages for which only a limited amount of Linked Data exist. The Semantic Quran dataset consists of all chapters of the Quran in 42 different languages including rare languages such as Divehi, Amazigh and Amharic. The data were extracted from two semi-structured sources: the data from the Tanzil project and the Quranic Arabic Corpus (cf. section 4). We designed an ontology for representing multilingual data extracted from sources consisting of different translations to the Quran book

[1] From the 315 datasets analyzed by the LodStats framework (`http://stats.lod2.eu`), 128 datasets provide English and American labels. French, the second most popular language in the LOD Cloud, is used in only 15 (approximately 4.8%) of the datasets. Most other languages occur in one dataset.

as well as numbered chapters and verses. In addition to providing aligned translations for each verse, we provide morpho-syntactic information on each of the original Arabic terms utilized across the dataset. Moreover, we interlinked the dataset to three versions of *Wiktionary* as well as *DBpedia* and ensured therewith that our dataset abides by all the four Linked Data principles[2].

In the following, we present the data sources that we used for the extraction (section 2). Thereafter, we give an overview of the ontology that underlies our dataset (section 3). Section 4 depicts the extraction process that led to the population of our ontology. We present our approach to interlinking the Semantic Quran and Wiktionary in section 5. Finally, we present several usage scenarios for the dataset at hand (section 6).

2. Data Sources

Two web resources were used as raw data sources for our dataset. The first web resource is the data generated by the *Tanzil Project*[3], which consists of the original verses in Arabic as well as 42 manual translations of the entire book. Our second web resource, *the Quranic Arabic Corpus*[4], was used to obtain morpho-syntactic information on each of the words contained in the Arabic version of the Quran.

2.1. Tanzil Project

The Tanzil Project[5] was motivated by inconsistencies across the different digital versions of the Quran. These were mainly due to missing/incorrect diacritics, Arabic text conversion problems, and missing encoding for some Arabic characters.

Tanzil was launched in early 2007 with the aim of producing a curated unicode version of the Arabic Quran text that can serve as a reliable standard text source on the web. To achieve this goal, Tanzil team subsequently developed a three-step data quality assurance pipeline; which consisted of (1) an automatic text extraction of Arabic Quran text, (2) a rule-based verification of the Arabic Quran text against a set of grammatical and recitation rules and (3) a final manual verification by group of experts.

The result of this process was a set of datasets that were made available in several versions and formats[6] In addition to the original Arabic sources, Tanzil provides translations of the Quran in 42 different languages by different translators[7]. We manually selected one translation per language for the extraction process[8]. Note that all Tanzil datasets are distributed under the terms of Creative Commons Attribution 3.0 License[9]. For the sake of completeness, we also added the content of descriptive books *"Tafseer Al Galalen"* by *"Jalal ad-Din al-Mahalli and Jalal ad-Din as-Suyuti"* and *"Tafser Al Moyaser"* by *"King Fahad Quran Complex"*.

2.2. The Quranic Arabic Corpus Project

The Quranic Arabic Corpus is an open source project, which provides Arabic annotated linguistic resources which shows the Arabic grammar, syntax and morphology for each word in the Quran. This is due to the rich morphological structure of the Arabic Language. In particular, a single word can encompass the semantics of entire english sentences. For instance the Arabic word *"faja'alnāhum"* can be translated into the entire English sentence "and we made them". The compact syntax of Arabis leads to that a single word being separable into distinct morphological segments. For example, *"faja'alnāhum"* can be subdivided into:

- *fa* – a prefixed conjunction ("and")
- *ja'al* – the stem, a perfect past tense verb ("made") inflected as first person masculine plural
- *nā* – a suffixed subject pronoun ("we")
- *hum* – a suffixed object pronoun ("them")

A Resource Description Framework (RDF) and Natural Language Processing Interchange Format (NIF)[4] representation of this rich morphology promises to further the development of integrated NLP pipelines for processing Arabic. In addition, given that this corpus was curated manually by experts, it promises to improve the evaluation of integrated NLP frameworks. We thus decided to integrate this data with the translation data available in the Tanzil datasets. Note that we

[2] http://www.w3.org/DesignIssues/LinkedData.html

[3] http://tanzil.net/

[4] http://corpus.quran.com

[5] http://tanzil.net/wiki/Tanzil_Project

[6] For more details on available formats and datasets, please see http://tanzil.net/download/.

[7] http://tanzil.net/trans/.

[8] The list of translations used can be found at http://goo.gl/s5RuI

[9] http://creativecommons.org/licenses/by/3.0/

used the Quranic Arabic Corpus Version 0.4[10] in its delimited text file version under the *"GNU General Public License"*.[11] Also note that the Quranic Arabic Corpus was developed after the verified Arabic text distributed by the Tanzil project and can thus be aligned with it.

3. Ontology

To represent the data as RDF, we developed a general-purpose linguistic vocabulary. The vocabulary[12] was specified with the aim of supporting datasets which display a hierarchical structure. It includes four basic classes: Chapter, Verse, Word and LexicalItem.

- *Chapter*: The chapter class provides the name of chapters in different languages and localization data such as chapter index and order. Additionally, the chapter class provides metadata such as the number of verses in a chapter, revelation place and some provenance information. Finally, the chapter class provides inter-class linking properties to link it to different verses contained in it. For example each chapter provides a `dcterms:tableOfContents` for each of its verses in the form `qrn:quran<chapter>-<verse>`.
- *Verse*: The verse class contains the verse text in different languages as well as numerous localization data such as verse index and related chapter index. Additionally, this class provides related verse data such as different verse descriptions and provenance information. Finally, it contains inter-class linking properties to link the verse in both directions; the chapter of the verse and the whole words contained in such a verse.
- *Word*: This class encompasses the verse next level of granularity and contains the word text in different languages as well as numerous localization data such as related verse and chapter verse indexes. Additionally, the word class provides related word provenance information and some inter-class linking properties to link it to chapter and verse of such a word.
- *LexicalItem*: Currently, this class provides morphological data on the Arabic words only. This class is compatible with the *GOLD linguistic ontology*[13] [2]. It provides variations of GOLD linguistic properties for each lexical item, such as acoustic, root, part of speech, gender-number-person properties. All the objects of the previously mentioned properties are URIs from the *OLIA Arabic Linguistic ontology*[14] . Finally, the lexical item class provides related lexical items provenance information, and also some inter-class linking properties to link it to chapter, verse, word of such a lexical item.

A UML class diagram of the four basic ontology classes of the Semantic Quran Dataset with inter-class internal relations is shown in Figure 1.

4. Extraction Process

The original Tanzil Arabic Quran data and translations are published in various formats. For simplicity purpose, delimited text files were selected as the basis for the RDF extraction. The format of the delimited file was `chapterIndex|verse|verseText`. For example, the first verse of the first chapter of the English translation of the Quran is `1|1|In the Name of Allah, the Most Beneficent, the Most Merciful.` On the other hand, the Quranic Arabic corpus is available as tab-separated text file of the form `"LOCATION FORM TAG FEATURES"`:

- The `LOCATION` field consists of 4-part numbering scheme of the form (Chapter : Verse : Word : Segment). For example, the first segment of the first word of the first verse of the first chapter has the form (1:1:1:1).
- The `FORM` field contains the text of the current segment in the Extended *Buckwalter transliteration*[15].
- The `TAG` field contains the part-of-speech tag for the current segment.
- The `FEATURES` field contains a complete morphological analysis of the current segment such as root, case and person-number-gender properties.

[10] `http://corpus.quran.com/download/`
[11] `http://www.gnu.org/licenses/gpl.html`
[12] `http://www.mlode.nlp2rdf.org/quranvocab#`
[13] `http://linguistics-ontology.org/`
[14] `http://nachhalt.sfb632.uni-potsdam.de/owl/`
[15] The Buckwalter transliteration uses ASCII characters to represent the orthography of the Arabic language. For the conversion table, see `http://www.qamus.org/transliteration.htm`

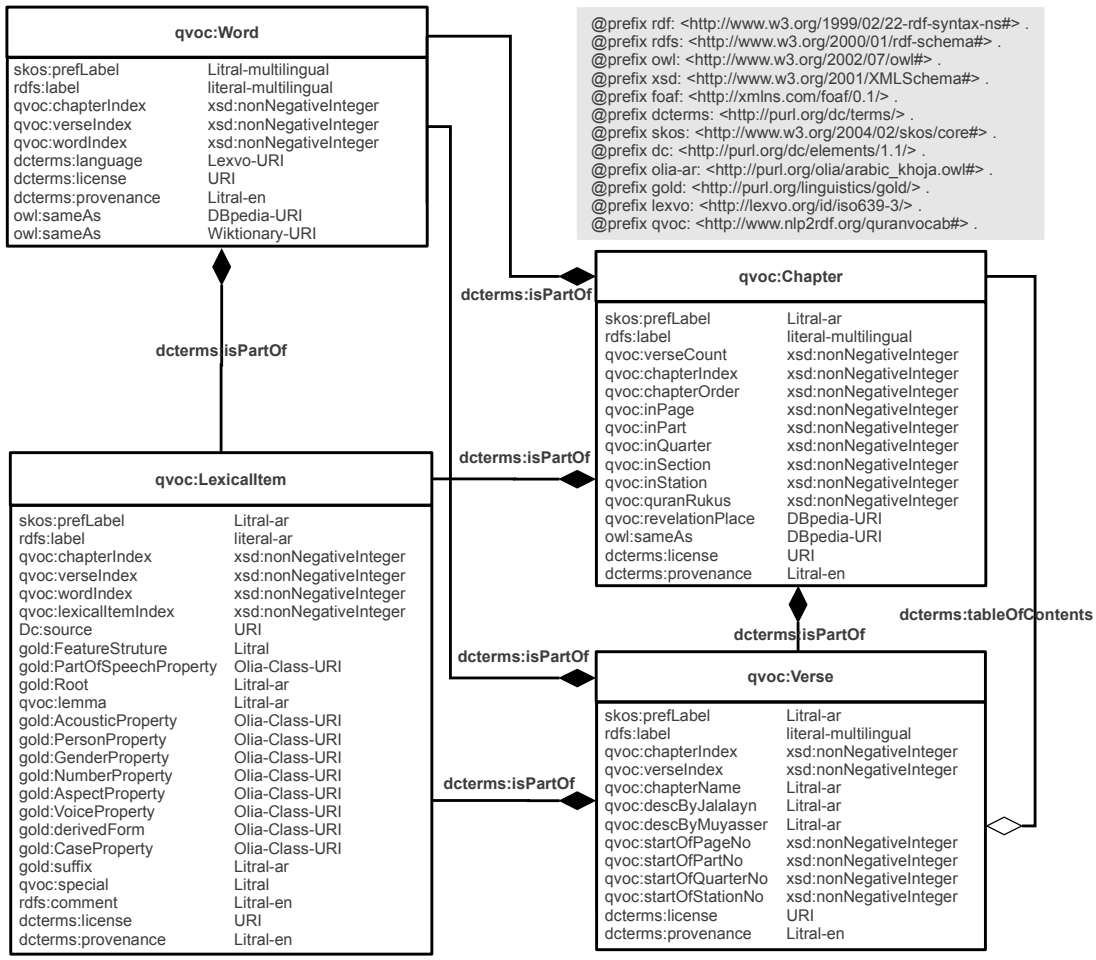

Fig. 1. UML class diagram of the Semantic Quran Ontology

Given the regular syntax used in the text file corpus at hand, we were able to carry out a one-to-one mapping of each fragment of the input text file to resources, properties or data types as explicated in the ontology shown in Figure 1. We relied on the *Apache Jena* Framework for the conversion. Note that for the sake of reuse, we created a converter class for each of the languages at our disposition. Therein, we were able to model the specifics of each language at hand. In addition, we developed a parser for the raw original Quranic Arabic Corpus dataset. The part-of-speech information and morphological analysis of each segment of the Arabic Quranic Corpus were extracted by using this parser and integrated to the words found in the Tanzil dataset. The merged data is now available in the Turtle RDF format. In order to simplify the interoperability of the generated dataset, we followed the specifications of the NIF. Currently, the original Arabic and four different translations of Quran (Arabic, English, German, French and Russian) abide by the NIF formalization. Details of Semantic Quran dataset CKAN entry, SPARQL endpoint, version and license are listed in Table 1

Table 1
Technical details of the Quran RDF dataset.

Name	SemanticQuran
URL	`http://www.mlode.nlp2rdf.org/semanticquran/`
Sparql Endpoint	`http://mlode.nlp2rdf.org/sparql`
Ver. Date	29.11.2012
Ver. No	0.1
Licence	Attribution-NonCommercial-ShareAlike 3.0 Unported (CC BY-NC-SA 3.0)
CKAN	SemanticQuran

5. Linking

We aim to link our dataset with as many data sources as possible to ensure maximal reusability and integrability in existing platforms. We have generated links to 3 versions of the RDF representation of Wiktionary as well as to DBpedia. All links were generated by using the LIMES framework [5]. The link specification used was essentially governed by fragments similar to that shown in Listing 1. The basic intuition behind this specification is to link words that are in a given language in our dataset to words in the same language with have exactly the same label. We provide 7617 links to the English version of DBpedia, which in turn is linked to non-English versions of DBpedia. In addition, we generated 7809 links to the English, 9856 to the French and 1453 to the Germany Wiktionary. Links to further versions of DBpedia and Wiktionary will be added in the future.

```
<SOURCE>
<ID>quran</ID>
<ENDPOINT>http://www.mlode.nlp2rdf.org/qurankarem/</ENDPOINT>
<VAR>?x</VAR>
<PAGESIZE>-1</PAGESIZE>
<RESTRICTION>?x a qvoc:Word</RESTRICTION>
<PROPERTY>rdfs:label AS lowercase->nolang
  </PROPERTY>
</SOURCE>
<TARGET>
<ID>wiktionary</ID>
<ENDPOINT>http://wiktionary.dbpedia.org/sparql
  </ENDPOINT>
<VAR>?y</VAR>
<PAGESIZE>-1</PAGESIZE>
<RESTRICTION>?y rdf:type lemon:LexicalEntry
  </RESTRICTION>
<RESTRICTION>FILTER langMatches( lang(?v0), "en" )
  </RESTRICTION>
<PROPERTY>rdfs:label AS lowercase->nolang
  </PROPERTY>
</TARGET>
<METRIC>trigrams(x.rdfs:label,y.rdfs:label)</METRIC>
```

Listing 1: Fragment of the link specification used to link the Semantic Quran and the English-language version of Wiktionary.

6. Use-Cases

The availability of a multilingual parallel corpus in RDF promises to facilitate a large number of NLP applications. In this section, we outline selected application scenarios and use-cases for the Semantic Quran data.

Data Retrieval. Since the Quran contain a lot of data concerning places, people and different events, multilingual sentences concerning such information can be easily retrieved from the dataset. The aligned multilingual representation allows searching for the same entities across different languages. For example, Listing 2 shows a SPARQL query which allows retrieving Arabic, English and German translations of verses which contain *"Moses"*.

```
SELECT ?verseIndex ?chapterName ?verseTextAr ?verseTextEn ?verseTextGr
WHERE{
?word rdfs:label "moses"@en;
    dcterms:isPartOf ?verse.
?verse a qvoc:Verse;
     skos:prefLabel ?verseTextAr;
     qvoc:verseIndex ?verseIndex;
     dcterms:isPartOf ?chapter;
     rdfs:label ?verseTextEn.
        FILTER ( lang(?verseTextEn) = "en" )
     rdfs:label ?verseTextGr.
        FILTER ( lang(?verseTextEn) = "de" )
?chapter skos:prefLabel ?chapterName.
}
```

Listing 2: Verses that contains moses in (i) Arabic (ii) English and (iii) German.

Arabic Linguistics. The RDF representation of Arabic morphology and syntax promises to facilitate the retrieval of relevant sub-corpora for researchers in linguistics. For example, Listing 3 provides an example of a SPARQL query witch retrieves all Arabic prepositions as well as an example statement for each of them.

```
SELECT ?preposition sample (?verseTextAr)
WHERE{
?word gold:PartOfSpeechProperty olia-ar:Preposition;
     skos:prefLabel ?preposition;
     dcterms:isPartOf ?verse.
?verse a qvoc:Verse;
     skos:prefLabel ?verseTextAr.
}GROUP BY ?preposition
```

Listing 3: List all the Arabic prepositions with example statement for each.

Another example is provided by Listing 4, which shows a list of different part-of-speech variations of one Arabic root of the word read *"ktb"*; note that in this example we use the Arabic root *"ktb"* witten in The Buckwalter transliteration.

```
SELECT DISTINCT ?wordText ?pos
WHERE{
?wordPart a qvoc:LexicalItem ;
        gold:Root "ktb";
        dcterms:isPartOf ?word;
        gold:PartOfSpeechProperty ?pos.
?word a qvoc:Word;
      skos:prefLabel ?wordText.
}
```

Listing 4: List of different part of speech variations of one Arabic root of the word read *"ktb"*.

interoperability using NIF. Using the interoperability capabilities provided by NIF, it is easy to query all occurrences of a certain text segment without using the verse, chapter, word, or lexical item indexes. For instance, Listing 5 lists all the occurrences of *"Moses"* with no need to have an extra index.

```
select ?verse {
?s str:occursIn ?verse.
?allah str:referenceContext ?s.
?allah str:anchorOf "Moses"@en
}
```

Listing 5: List of all occurrences of "Moses" using NIF

Information Aggregation. The interlinking of the Quran dataset with other RDF data sources provides a considerable amount of added value to the dataset. For example, the interlinking with Wiktionary can be used as in Listing 6 to get the different senses for each of the English words contained in the first verse of the first chapter *"qrn:quran1-1"*.

```
SELECT ?wordIndex ?WordText ?sense
WHERE{
?word a qvoc:Word;
    dcterms:language lexvo:eng ;
    dcterms:isPartOf qrn:quran1-1;
    qvoc:wordIndex ?wordIndex;
    rdfs:label ?WordText;
        FILTER ( lang(?verseTextEn) = "en" )
    owl:sameAs ?wiktionaryWord.
?wiktionaryWord lemon:sense ?sense
}
```

Listing 6: List of all senses of all English words of the first verse of the first chapter "qrn:quran1-1".

7. Conclusion and Future Work

In this work, we presented the Semantic Quran, an integrated parallel RDF dataset in 42 languages. This multilingual dataset aims to increase the availability of multilingual data in LOD and to further the development of NLP tools for languages that are still under represented, if not absent, from the LOD cloud. Thanks to its RDF representation, our dataset ensures a high degree of interoperability with other datasets. For example, it provides 26735 links overall to Wiktionary and DBpedia. As demonstrated by our use-cases, the dataset and the links it contains promise to facilitate research on multilingual applications. Moreover, the availability of such a large number of languages in the dataset provides opportunities for linking across the monolingual datasets on the LOD Cloud and thus perform various types of large-scale analyses.

As future work, we aim to extend the data by providing interlinks to the upcoming versions of Wiktionary. Additionally, we will interlink the Semantic Quran dataset with many of the publicly available multilingual *Wordnets*. We already provided NIF for the five languages Arabic, English, French, German and Russian. We will extend the NIF content of the dataset to the remaining 38 languages.

References

[1] Sören Auer, Jens Lehmann, and Axel-Cyrille Ngonga Ngomo. Introduction to linked data and its lifecycle on the web. In *Reasoning Web*, pages 1–75, 2011.

[2] Scott Farrar and Terry Langendoen. A linguistic ontology for the semantic web. *GLOT INTERNATIONAL*, 7, 2003.

[3] Sebastian Hellmann. The semantic gap of formalized meaning. In *ESWC (2)*, volume 6089 of *Lecture Notes in Computer Science*, pages 462–466. Springer, 2010.

[4] Sebastian Hellmann, Jens Lehmann, and Sören Auer. Linked-data aware uri schemes for referencing text fragments. In *EKAW 2012*, Lecture Notes in Computer Science (LNCS) 7603. Springer, 2012.

[5] Axel-Cyrille Ngonga Ngomo. On link discovery using a hybrid approach. *Journal on Data Semantics*, 1:203 – 217, December 2012.

[6] H. Somers. Machine translation and minority languages. *Translating and the Computer*, pages 13–13, 1997.

[7] Christina Unger, Lorenz Bühmann, Jens Lehmann, Axel-Cyrille Ngonga Ngomo, Daniel Gerber, and Philipp Cimiano. Sparql template-based question answering. In *Proceedings of WWW*, 2012.