Using WikiData as a multi-lingual multi-dialectal dictionary for Arabic dialects

Houcemeddine Turki, University of Sfax
Denny Vrandečić, Google
Helmi Hamdi, University of Sherbrooke
Imed Adel, Lycée secondaire Sbikha 1979
Using WikiData as a multi-lingual multi-dialectal dictionary for Arabic dialects

Conference Paper · October 2017
DOI: 10.1109/AICCSA.2017.115

CITATIONS
0

READS
171

4 authors, including:

Houcemeddine Turki
University of Sfax
30 PUBLICATIONS  4 CITATIONS

Imed Adel
Lycée secondaire Sbikha 1979, Tunisia
11 PUBLICATIONS  3 CITATIONS

Some of the authors of this publication are also working on these related projects:

- Studying the evolution of Interdisciplinary Borrowing in Economics View project
- Reformed Modern Standard Arabic View project

All content following this page was uploaded by Houcemeddine Turki on 13 November 2017.
The user has requested enhancement of the downloaded file.
Using WikiData as a multi-lingual multi-dialectal dictionary for Arabic dialects

Houcemeddine Turki
Faculty of Medicine of Sfax
University of Sfax
Sfax, Tunisia
turkiabdelwaheb@hotmail.fr

Denny Vrandečić
Google
San Francisco, USA
vrandecic@google.com

Helmi Hamdi
Faculté des Sciences
Université de Sherbrooke
Sherbrooke, Québec, Canada
helmoony@hotmail.com

Imed Adel
1979 Sbikha High School
Sbikha, Kairouan, Tunisia
geekemad@gmail.com

Abstract—Since 2012, Wikidata has been developed as a freely accessible and community generated knowledge database that represents not only the name of each item (tree, famous person...) in all available languages but also to define links (like "component of", "born in", "instance of"...) between items. Nowadays, the output of Wikidata has become one of the greatest semantic web data in the world and has been proved to be useful to solve many currently existing problems in Computational Linguistics, in Medicine, and in many other fields. In this research work, we propose to convert Wikidata into a multi-lingual multi-dialectal dictionary for Arabic dialects and we explain how Wikidata as a multi-lingual multi-dialectal dictionary for Arabic dialects can be later used for the natural language processing of the varieties of Arabic by computational linguists and computer scientists.

Keywords—WikiData, Multidialectal dictionary, Multilingual dictionary, Arabic dialects

I. INTRODUCTION

Nowadays, the sociolinguistic situation of the Arab world is quite complicated [1]. In fact, farther than the spread of Modern Standard Arabic and several European languages in the Middle East and North Africa, the Arab world is characterized by the existence of an important number of Arabic dialects [1], [2]. These dialects are not all mutually intelligible and each one of them has its morphological and phonological and even semantic and lexical particularities [1, 2]. That is why many studies and linguistic works about these Arabic dialects have been conducted during the 20th and the 21st centuries [3, 4, 5, 6]. These efforts did not only consist of the creation of simple morphological [3, 4, 5], pragmatic [7, 8] and phonological descriptions [9, 10, 11] of the Arabic dialects. These efforts also involved the development of high-scale English - Dialectal Arabic dictionaries for each diverse Arabic dialect [12, 13, 14, 15]. This led to the creation of the first multi-dialectal multi-lingual dictionary for Arabic dialects: Tharwa [16, 17]. This dictionary extracts its data from several bilingual Dialectal Arabic – English electronic dictionaries and gives the synonyms of the Egyptian Arabic words in Modern Standard Arabic, in Levantine Arabic, and in English [16, 17].

In this research work and as a continuation of the work about the Tharwa dictionary, we will discuss if and how we can use WikiData, an online multilingual freely available knowledge database, as a more performant multidialectal multi-lingual multi-dialectal dictionary for all the currently existing Arabic dialects. We will begin our discussion by giving an overview about the principles of work and the structure of Wikidata (Part II). Then, we will explain how it is possible to convert Wikidata into a multi-lingual multi-dialectal dictionary for Arabic dialects (Part III). Finally, we will describe how Wikidata (as a multilingual multidialectal dictionary for Arabic dialects) can be used by Computational linguists and Computer scientists in the Natural Language Processing of the varieties of Arabic (Part IV).

II. WIKIDATA: AN OVERVIEW

Developed by Wikimedia Foundation, Google Inc. and Wikimedia Deutschland e.V. since 2012 as a proliferation of DBPedia and Freebase projects and accessible online in https://www.wikidata.org, Wikidata is an open and collaborative knowledge base, originally introduced to support Wikipedia [18, 19] but has since grown widely in use [20].

The principle of Wikidata is simple. Users retrieve items (scientific phenomena, people, objects...) from the Wikimedia wikis like Wikipedia and Wiktionary and from other databases like ICPC, Patientplus, ICD, and MeSH among others [20]. Then, items having relationship between them are linked together using properties as shown in Fig. 1 [19].

![Fig. 1. An example of how items and their data are interconnected in WikiData (A number beginning by Q like Q42 [Douglas Adams] is assigned to define each item, A number beginning by P like P31 [Instance of] is assigned to define the property of each link, Credits: John Erling Blad - CC BY-SA 3.0 License)](image-url)
As the links between items are added to Wikidata by human contributors and to let such links verifiable, existing references proving the accuracy of the links are added to Wikidata as well as shown in red in Fig. 2 [20].

Later, to let the data provided by Wikidata readable by users from all over the world, a label, a list of aliases and a description written in each language having an ISO 639-3 code are assigned to each item or property available in Wikidata as shown in Fig. 3 [20].

Table I. List of ISO 639-3 codes for the Arabic dialects [26]

<table>
<thead>
<tr>
<th>Arabic dialect</th>
<th>ISO 639-3 code</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tunisian</td>
<td>AEB-ARAB</td>
</tr>
<tr>
<td>Algerian</td>
<td>ARQ</td>
</tr>
<tr>
<td>Moroccan</td>
<td>ARY</td>
</tr>
<tr>
<td>Egyptian</td>
<td>ARZ</td>
</tr>
<tr>
<td>North Levantine</td>
<td>APC</td>
</tr>
<tr>
<td>South Levantine</td>
<td>AJP</td>
</tr>
<tr>
<td>Cypriot</td>
<td>ACY</td>
</tr>
<tr>
<td>Libyan</td>
<td>AYL</td>
</tr>
<tr>
<td>Sudanese</td>
<td>APD</td>
</tr>
<tr>
<td>Saharian</td>
<td>AAO</td>
</tr>
<tr>
<td>Hassaniya</td>
<td>MEY</td>
</tr>
</tbody>
</table>

As Arabic dialects do not generally have any officially recognized regulatory institution or any official standard orthography guidelines (although the important efforts that were done in this context) and as all labels in Wikidata in an Arabic dialect should be written using a unique normalized orthography so that these labels would not be difficultly read [27], we chose to write the labels in Wikidata in Arabic script following the CODA guidelines that were originally defined by Habash et al. in 2012 [27] and as completed by the general and particular CODA extensions defined for Tunisian [28], Wikipedia. As well, in Wikidata, the lexical networks of the Arabic dialects are parallel to the ones of the other languages available in Wikidata as the users of each languages just added labels to the same network of items existing in Wikidata [19]. Furthermore, Wikidata is also easier and cheaper to access online than the Arabic dialects’ WordNet and Treebank [20]. In fact, the data available in Wikidata can be freely found online using the user-friendly interface available in www.wikidata.org and shown in Fig. 2 [20]. As well, wikidata is based on widely used web standards, such as XML, JSON and RDF for representing its content, and SPARQL as a query language [19, 20]. This allows the easy use of the data with any standard compliant tools by mobile apps, computer software, or online interfaces [20].

III. METHOD OF LETTING WIKIDATA A MULTI-LINGUAL MULTI-DIALECTAL DICTIONARY FOR ARABIC DIALECTS

A. Principle and Orthography guidelines

As shown before, to use Wikidata as a multi-lingual multdialectal dictionary for Arabic dialects, contributors should only assign labels for all the Wikidata items and properties in all Arabic dialects through the user-friendly interface accessible in www.wikidata.org [20]. This method is quite the same as the one used for the automatic creation of bilingual dictionaries for Finno-Ugric languages and it only supports Arabic dialects having ISO 639-3 codes shown in Table I [25].

As Arabic dialects do not generally have any officially recognized regulatory institution or any official standard orthography guidelines (although the important efforts that were done in this context) and as all labels in Wikidata in an Arabic dialect should be written using a unique normalized orthography so that these labels would not be difficultly read [27], we chose to write the labels in Wikidata in Arabic script following the CODA guidelines that were originally defined by Habash et al. in 2012 [27] and as completed by the general and particular CODA extensions defined for Tunisian [28],...
Algerian [29], Egyptian [30, 31], Maghrebi [32], Levantine (particularly Palestinian) [33], and Gulf [34] dialects. Created as a development of [35] and [36], the principle of CODA guidelines is to transcribe texts in Arabic dialects using the orthography rules of Modern Standard Arabic and by taking into consideration the similarities between Modern Standard Arabic and Arabic Dialects [27]. Further information about this orthography guidelines can be found in [27-34].

B. Adding or modifying labels, descriptions, and aliases to a Wikidata item or property

To add, adjust or remove the labels, descriptions, and aliases to the page of a Wikidata item or property, the user should simply create a Wikimedia account or log in to Wikidata as shown in Fig. 4.

![Fig. 4. How to create a Wikimedia account [shown here in red] and how to log in to Wikidata [shown here in green] (Sources: https://www.wikidata.org/wiki/Wikidata:Main_Page, https://www.wikidata.org/w/index.php?title=Special:CreateAccount, and https://www.wikidata.org/w/index.php?title=Special:UserLogin, Credits: Houcemeddine Turki - CC BY-SA 3.0 License)](image)

Then, he should choose the language of the label that will be created or adjusted as shown in Fig. 5.

![Fig. 5. How to choose the language of the label that will be created, adjusted, or removed (Source: https://www.wikidata.org/wiki/Wikidata:Main_Page, Credits Houcemeddine Turki - CC BY-SA 3.0 License)](image)

After that, he should access to the page of the concerned Wikidata item or property (like https://www.wikidata.org/wiki/Q56426), click on Edit above the LabelBox, add or adjust the label, description and aliases of the Wikidata item or property in the chosen Arabic dialect, and finally click on Save as shown in Fig. 6.

![Fig. 6. How to add or adjust the label in the chosen Arabic dialect through the page (Source: https://www.wikidata.org/wiki/Q56426, Credits: Houcemeddine Turki - CC BY-SA 3.0 License)](image)

C. Using Wikidata as a multi-lingual multi-dialectal dictionary for Arabic dialects

The use of Wikidata as a multi-lingual multi-dialectal dictionary for Arabic dialects is simple.

If the user knows the name of the searched Wikidata item or property in an initial Arabic dialect (like Tunisian and Algerian) or an initial foreign language (like French and English), he just has to log in as shown in Fig. 4, choose the initial language or Arabic dialect as shown in Fig. 5, and finally write the name of the Wikidata property or item in the chosen language or Arabic dialect in the search box in the main page of Wikidata and click on one of the option automatically appeared as shown in Fig. 7.

![Fig. 7. How to search a Wikidata item or property by typing its label in a chosen language in the search box of the main page of Wikidata (Source: https://www.wikidata.org/wiki/Wikidata:Main_Page, Credits: Houcemeddine Turki - CC BY-SA 3.0 License)](image)

Consequently, you will access the page of the searched Wikidata item or property. You can find the name of the Wikidata item or property in the target language or Arabic dialect in the Labelbox of that page as shown in Fig. 3.
Else, the user can search for the name of a Wikidata item in a given Arabic dialect using the Wikidata query service if he knows the property of each relationship between the item and other Wikidata items. Available in https://query.wikidata.org, this query service runs over an RDF representation of Wikidata that is indexed in the Blazegraph SPARQL store [37]. Using this alternative service is simple. The user should write the SPARQL code of the requested query (like the one shown in Fig. 8) and then click on Run (as shown in Fig. 9).

Consequently, Wikidata can be used to find the false friends between Arabic dialects (like “ثِقَة” meaning “hazelnut” in Modern Standard Arabic and “pine nut” in Tunisian) as well as the near minimal and minimal pairs in the Arabic dialects (e.g.: “فرفط” and “رففط” are not minimal pairs in Tunisian because they both mean butterfly. However, “رففط” and “رففط” are near minimal pairs in Tunisian because they respectively mean butterfly and wasp). As well, Wikidata can be used to find technical words in Arabic dialects. This will help solving the problem of the lack of intelligibility of the spoken discourse of professionals (like physicians) to clients (like patients) because professionals will not be obliged to use code-switching to French or English to explain their works [38].

B. Indirect functions

Wikidata has many indirect functions that have been reported in previous researches:

- **Machine Translation**: Wikidata can be used as one of the reference databases of machine translation systems as it is a multi-lingual multi-dialectal dictionary for Arabic dialects providing, when completed, the name of every item in all the main Arabic dialects. Unfortunately, before Wikidata, such dictionaries did not exist and this is what explains the current poor accuracy of the machine translation of Arabic dialects [39]. Furthermore, Wikidata can give a needed data for the accuracy of a machine translation like the gender of a famous person as shown in Fig. 10 [40]. In fact, Rabinovich et al. proved in 2017 that using Wikidata to recognize the gender of the people that are evocated in the input of machine translation (MT) systems helps to solve all gender-related deficiencies of the work of such systems and stipulates that retrieving other types of data related to specified items in the input of MT systems can solve other problems causing the lack of accuracy of these systems [40].

**Example in Wikidata**

- Camille Lacourt is a French swimmer.
- Camille can be the first name of a man or a woman
- Swimmer is used to call a person who swims whatever his/her gender is.
- No data in the sentence about the gender of Camille Lacourt
- Querying Wikidata
  - Gender (P21): Male
- (Tunisian) "ل又好 هو عُازم فرنساوي" كامي لاقور هو عَازم فرنساوي

**Morphological Analysis and Generation**: Wikidata can be used to automatically find and generate the morphological rules of a supported language [41]. In fact, as links between any two Wikidata items can be easily found through https://www.wikidata.org, statistically analyzing all the sentences containing the two items in social networks (like LinkedIn) and

---

IV. USEFUL FUNCTIONS OF WIKIDATA AS A MULTI-LINGUAL MULTI-DIALECTAL DICTIONARY FOR ARABIC DIALECTS

As a multi-lingual multi-dialectal dictionary for Arabic dialects, Wikidata has important direct functions. These functions can be accomplished through the simple and direct use of Wikidata. Furthermore, Wikidata also has indirect functions. These functions are done using software, apps, or online interface that use Wikidata as one of their database of reference. In this section, we try to explain most of the possible direct and indirect possible functions of Wikidata to give scientists an idea about how they can use this free and online database to ameliorate the accuracy of their ANLP research projects.

A. Direct functions

As shown before and just like Tharwa dictionary [16], Wikidata can be used to find the name of an item in a given Arabic dialect if the user knows its name in any other Arabic dialect or language, to find the semantic relationship between two terms in an Arabic dialect (like “synonym of”, “opposite to”, or “component of”…), or to find a name of an item in an Arabic dialect using the semantic relationships of that item with other ones.

---

![Fig. 8. The SPARQL query to retrieve the name of animals having “Panthera” as a parent taxon from Wikidata in Egyptian Arabic dialect](https://example.com/Sparql8.png)

**Fig. 8.** The SPARQL query to retrieve the name of animals having “Panthera” as a parent taxon from Wikidata in Egyptian Arabic dialect

![Fig. 9. How to use the Wikidata Query Service](https://example.com/Sparql9.png)

**Fig. 9.** How to use the Wikidata Query Service (Source: https://query.wikidata.org, Credits: Houcemeddine Turki - CC BY-SA 3.0 License)

---

![Fig. 10. An example of how Wikidata can be useful to ameliorate the accuracy of the translation process of a Machine translation system](https://example.com/Sparql10.png)

**Fig. 10.** An example of how Wikidata can be useful to ameliorate the accuracy of the translation process of a Machine translation system

---

Consequently, Wikidata can be used to find the false friends between Arabic dialects (like “وَّفَرَط” meaning “hazelnut” in Modern Standard Arabic and “pine nut” in Tunisian) as well as the near minimal and minimal pairs in the Arabic dialects (e.g.: “فرفط” and “رففط” are not minimal pairs in Tunisian because they both mean butterfly. However, “رففط” and “رففط” are near minimal pairs in Tunisian because they respectively mean butterfly and wasp). As well, Wikidata can be used to find technical words in Arabic dialects. This will help solving the problem of the lack of intelligibility of the spoken discourse of professionals (like physicians) to clients (like patients) because professionals will not be obliged to use code-switching to French or English to explain their works [38].

**B. Indirect functions**

Wikidata has many indirect functions that have been reported in previous researches:

- **Machine Translation**: Wikidata can be used as one of the reference databases of machine translation systems as it is a multi-lingual multi-dialectal dictionary for Arabic dialects providing, when completed, the name of every item in all the main Arabic dialects. Unfortunately, before Wikidata, such dictionaries did not exist and this is what explains the current poor accuracy of the machine translation of Arabic dialects [39]. Furthermore, Wikidata can give a needed data for the accuracy of a machine translation like the gender of a famous person as shown in Fig. 10 [40]. In fact, Rabinovich et al. proved in 2017 that using Wikidata to recognize the gender of the people that are evocated in the input of machine translation (MT) systems helps to solve all gender-related deficiencies of the work of such systems and stipulates that retrieving other types of data related to specified items in the input of MT systems can solve other problems causing the lack of accuracy of these systems [40].

**Example in Wikidata**

- Camille Lacourt is a French swimmer.
- Camille can be the first name of a man or a woman
- Swimmer is used to call a person who swims whatever his/her gender is.
- No data in the sentence about the gender of Camille Lacourt
- Querying Wikidata
  - Gender (P21): Male
- (Tunisian) "ل又好 هو عُازم فرنساوي" كامي لاقور هو عَازم فرنساوي

**Morphological Analysis and Generation**: Wikidata can be used to automatically find and generate the morphological rules of a supported language [41]. In fact, as links between any two Wikidata items can be easily found through https://www.wikidata.org, statistically analyzing all the sentences containing the two items in social networks (like LinkedIn) and

---

**IV. USEFUL FUNCTIONS OF WIKIDATA AS A MULTI-LINGUAL MULTI-DIALECTAL DICTIONARY FOR ARABIC DIALECTS**

As a multi-lingual multi-dialectal dictionary for Arabic dialects, Wikidata has important direct functions. These functions can be accomplished through the simple and direct use of Wikidata. Furthermore, Wikidata also has indirect functions. These functions are done using software, apps, or online interface that use Wikidata as one of their database of reference. In this section, we try to explain most of the possible direct and indirect possible functions of Wikidata to give scientists an idea about how they can use this free and online database to ameliorate the accuracy of their ANLP research projects.

**A. Direct functions**

As shown before and just like Tharwa dictionary [16], Wikidata can be used to find the name of an item in a given Arabic dialect if the user knows its name in any other Arabic dialect or language, to find the semantic relationship between two terms in an Arabic dialect (like “synonym of”, “opposite to”, or “component of”…), or to find a name of an item in an Arabic dialect using the semantic relationships of that item with other ones.

---

**Fig. 8.** The SPARQL query to retrieve the name of animals having “Panthera” as a parent taxon from Wikidata in Egyptian Arabic dialect

**Fig. 9.** How to use the Wikidata Query Service (Source: https://query.wikidata.org, Credits: Houcemeddine Turki - CC BY-SA 3.0 License)

---

**Fig. 10.** An example of how Wikidata can be useful to ameliorate the accuracy of the translation process of a Machine translation system

**Example in Wikidata**

- Camille Lacourt is a French swimmer.
- Camille can be the first name of a man or a woman
- Swimmer is used to call a person who swims whatever his/her gender is.
- No data in the sentence about the gender of Camille Lacourt
- Querying Wikidata
  - Gender (P21): Male
- (Tunisian) "ل又好 هو عُازم فرنساوي" كامي لاقور هو عَازم فرنساوي

**Morphological Analysis and Generation**: Wikidata can be used to automatically find and generate the morphological rules of a supported language [41]. In fact, as links between any two Wikidata items can be easily found through https://www.wikidata.org, statistically analyzing all the sentences containing the two items in social networks (like LinkedIn) and

---

440
Wikimedia wikis (The ones for Arabic dialects are still under development) will give the structure that is used in the analyzed language to express the studied link as shown in Fig. 11 [41].

![Fig. 11. An example of how Wikidata is used by morphological analyzers to find a structure used in the analyzed language for a specific function](image)

- **Part-of-speech tagging:** Because Wikidata provides data about each of the items it includes (like nature of the item, its year of construction...), it can be used to identify and mark (or tag) any item having a particular property in an input text including the ones written in Arabic dialects. That is why Klang et al. developed in 2016 Langforia, the first POS tagging set of tools having Wikidata as a reference database [42] and that is why Thompson used Wikidata as a reference database in the same year to verify the accuracy of his system that uses POS tags to identify locations in social network messages [43].

- **Identification of the general topic of texts and social network messages written in Arabic dialects:** As Wikidata includes technical terms, impolite words, and specific lexicons, the general topic of a given text written in an Arabic dialect can be automatically found by finding the key Wikidata items specified in the input text and then by finding out which is the specific lexicon that includes these words by analyzing the semantic network provided by Wikidata as shown in Fig. 12.

![Fig. 12. An example of how Wikidata is used by identify the general topic of a text written in an Arabic dialect](image)

Consequently, Wikidata can be used when completed to build algorithms to classify texts in Arabic dialects according to their topics, to analyze the sentiments of authors when writing their texts, and to detect impolite content like insults and bullying or harmful content like the ones encouraging criminal and terroristic behaviours within texts with a better accuracy than the one currently available (less than 85%) [44].

V. CONCLUSION AND FUTURE DIRECTIONS

In this research paper, we propose to convert Wikidata (a free, collaborative, and multilingual knowledge database available in https://www.wikidata.org) into an alternative and more performant multi-lingual multi-dialectal dictionary for Arabic dialects that can be easily completed, verified, adjusted and used by users and that can be also used to define links between semantic items (as data in Wikidata is organized in RDF format). We prove that Wikidata can do its direct functions (like finding words, synonyms, antonyms, false friends, and near minimal and minimal pairs) and its indirect functions (like Machine Translation, POS Tagging, and the identification of the general topics of texts for sentiment analysis or semantic classification of texts) more effectively than any other similar project such as the Arabic dialect Treebank, Tharwa dictionary, and the Arabic dialect WordNet (as the data contained in Wikidata is more voluminous). That is why we believe that Wikidata is an outstanding contribution to Computational Linguistics that can be used in the future as a reference database for other projects (apps, software, online interfaces...) for the NLP of Arabic dialects (Machine Translation between Arabic dialects) and we invite computer scientists to develop machine learning techniques to search the Internet and then to add labels to the undefined items in Arabic dialects or to add references to or remove unverified links between items.

VI. ACKNOWLEDGEMENTS

I should thank Derja Association, the regulatory institution of Tunisian, and Wikimedia TN User Group, the group of Tunisian contributors to Wikimedia projects, for their support to the project. I should also thank Prof. Mohamed Maamouri (Linguistic Data Consortium, University of Pennsylvania) and Prof. Adel Bouhoula (Université de Carthage) for their helpful comments and discussions.

REFERENCES

A. Thompson, "Using part-of-speech tags to identify presence of M. Abdul-Mageed, M. Diab and S. Kübler, "SAMAR: Subjectivity and View publication stats

[28]

[27]

[24]

[22]

[20]

[19]

[17]

[16]

[15]

[14]

[13]

[12]

[11]

[10]

[9]

[8]

[7]

[6]

[5]

[4]

[3]

[2]

[1]


M. Abdul-Mageed and M. T. Diab, "SANA: A Large Scale Multi-Genre, Multi-Dialect Lexicon for Arabic Subjectivity and Sentiment Analysis," in LREC, 2014.


V. Cavalli-Sforza, H. Suddiki, K. Bouzoubaa, L. Abouenour, M. Maamouri and E. Goshey, "Bootstrapping a wordnet for an arabic dialect from other wordnets and dictionary resources," in Computer systems and applications (acissa), 2013 acs international conference on, 2013.


M. Klang and P. Nugues, "Langforia: Language Pipelines for Arabic Subjectivity and Sentiment Analysis", in Proceedings of the Fifth International Conference on Language Resources and Evaluation, LREC’06, 2006.


