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](mailto:turkiabdelwaheb@hotmail.fr)

Denny Vrandečić

Google

San Francisco, USA [vrandecic@google.com](mailto:vrandecic@google.com)

Helmi Hamdi

Faculté des Sciences

Université de Sherbrooke

Sherbrooke, Québec, Canada

[helmoony@hotmail.com](mailto:helmoony@hotmail.com)

Imed Adel

1979 Sbikha High School

Sbikha, Kairouan, Tunisia

[geekemad@gmail.com](mailto: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 develop the use of Wikidata as a multi-lingual multi-dialectal dictionary for Arabic dialects and explain how this use of Wikidata can be useful for the amelioration of the accuracy and efficiency of the current Arabic natural language processing projects.

Keywords—WikiData, Multidialectal dictionary, Arabic dialects

# Introduction

Nowadays, the sociolinguistic situation of the Arab world is quite complicated [1]. In fact, further 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].

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 are not only limited to 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].

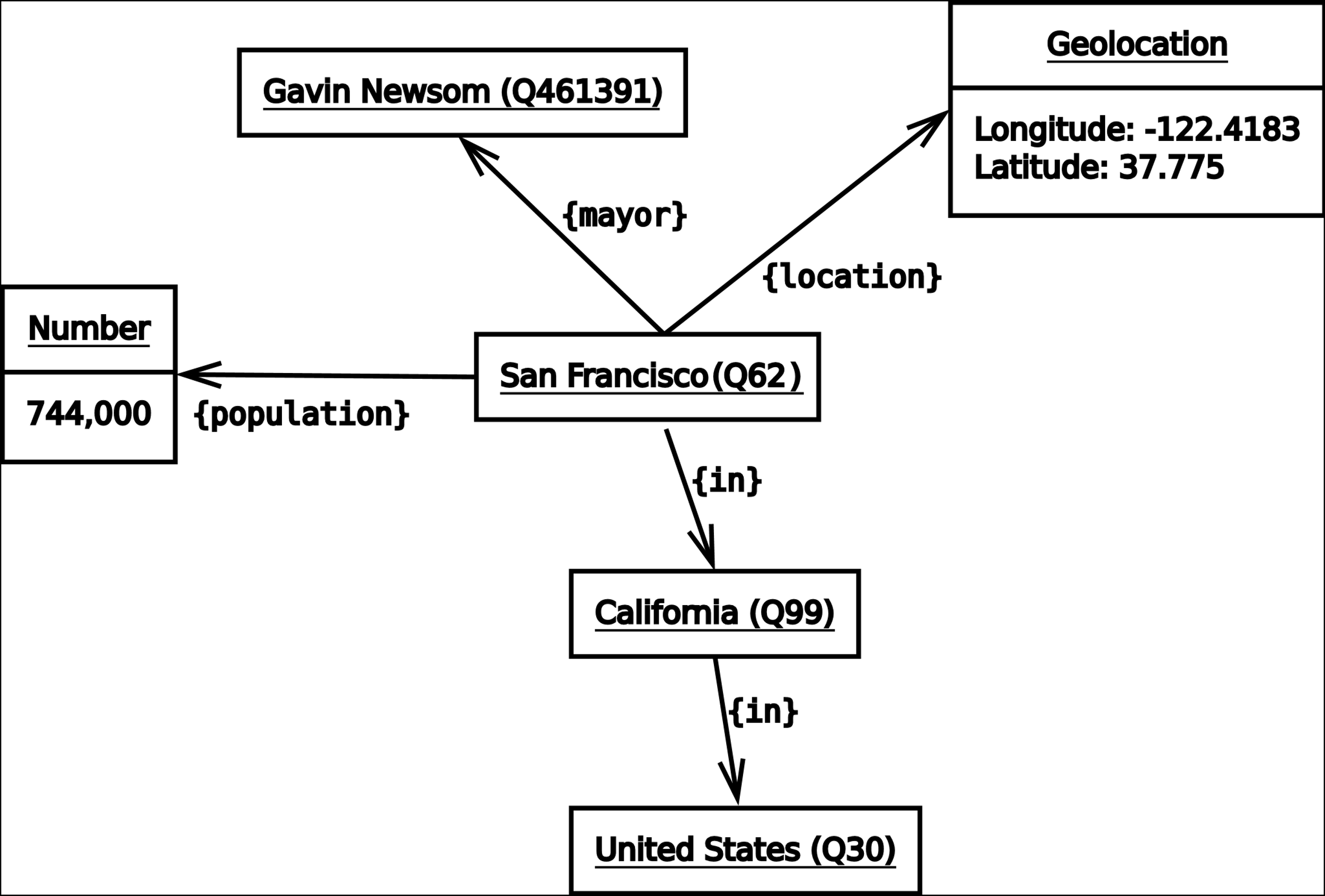
This led to the creation of the first multi-dialectal multi-lingual dictionary for Arabic dialects: Tharwa [16, 17]. This dictionary gives the synonyms of the Egyptian Arabic words in Modern Standard Arabic, in Levantine Arabic and in English [16, 17].

In this research work and as a continuation of the work about the Tharwa dictionary, we will discuss how to 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.

# WikiData: An overview

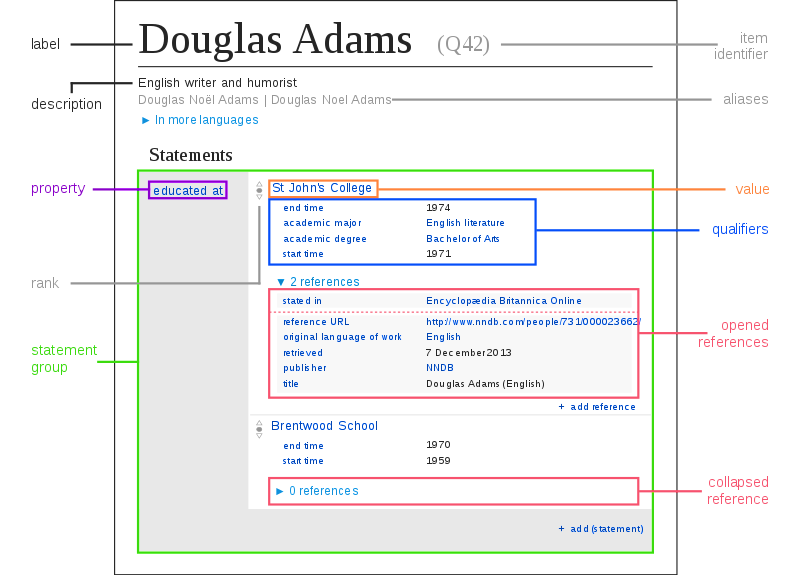
Developed by Wikimedia Foundation, Google Inc. and Wikimedia Deutschland e.V. since 2012 as a proliferation of DBPedia and Freebase projects and accessible online in <https://www.wikidata.org>, Wikidata is an open and collaborative knowledge base, originally introduced to support Wikipedia [18, 19] but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



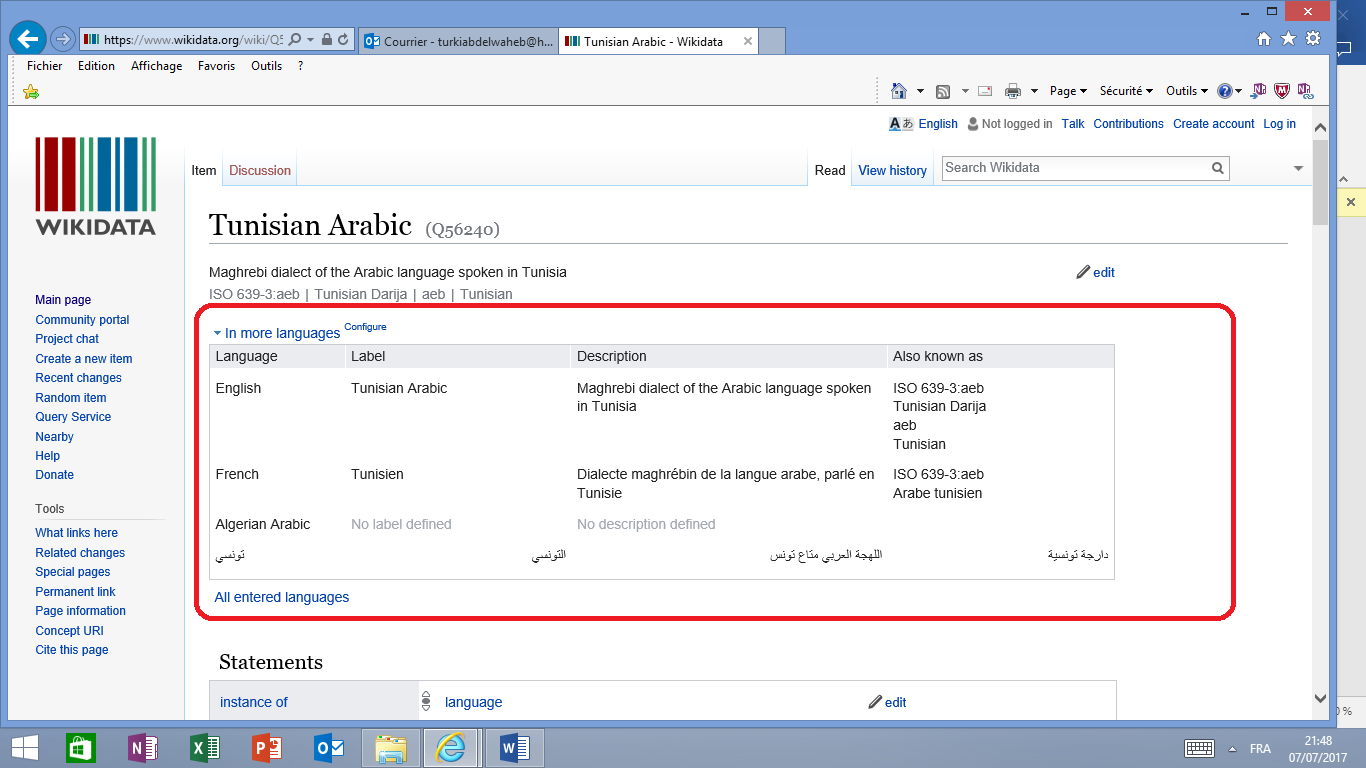
1. An example of how items and their data are interconnected in WikiData (A number beginning by Q like Q42 [Douglas Adams] is assigned to define each item, A number beginning by P like P31 [Instance of] is assigned to define the property of each link, Credits: John Erling Blad - CC BY-SA 3.0 License)

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].



1. Item accessed using the Graphical Interface of Wikidata (Source: <https://www.wikidata.org/wiki/Q42>, Credits: Charlie Kritschmar - CC BY-SA 3.0 License)

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].



1. The LabelBox in the page of the Wikidata item in the user-friendly interface of Wikidata featuring the labels, the description, and the aliases for the item in all the available languages shown in red (Source: <https://www.wikidata.org/wiki/Q56240>, Credits: Houcemeddine Turki - CC BY-SA 3.0 License)

Thanks to these contributions, a community generated semantic linked web data that reflects not only the relationship between common items (resp. common names and structures) but also the ones between items like celebrities, institutions and places (resp. proper names) was created in Wikidata [19]. So, the output of Wikidata is quite similar to the one of several projects aiming to the creation of a monolingual or multilingual semantic web that were applied before to Arabic dialects [19] like the WordNet project [21, 22] and the Arabic Treebank project [23, 24]. But, the volume of this lexical network is higher. In fact, as of January 2017, Wikidata has about 6,800 active editors, making it the third most active Wikimedia project before the German-language edition of Wikipedia. It contains more than 125 Million statements on about 20 Million items, and has a higher edit frequency than Wikipedia.

Furthermore, Wikidata is also easier to access online [20]. In fact, the data available in Wikidata can be found online using the user-friendly interface available in [www.wikidata.org](http://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].

# Method of Letting Wikidata a multi-lingual multi-dialectal dictionary for Arabic dialects

## Principle and Orthography guidelines

As shown before, to use Wikidata as a multi-lingual multi-dialectal 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](http://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].

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

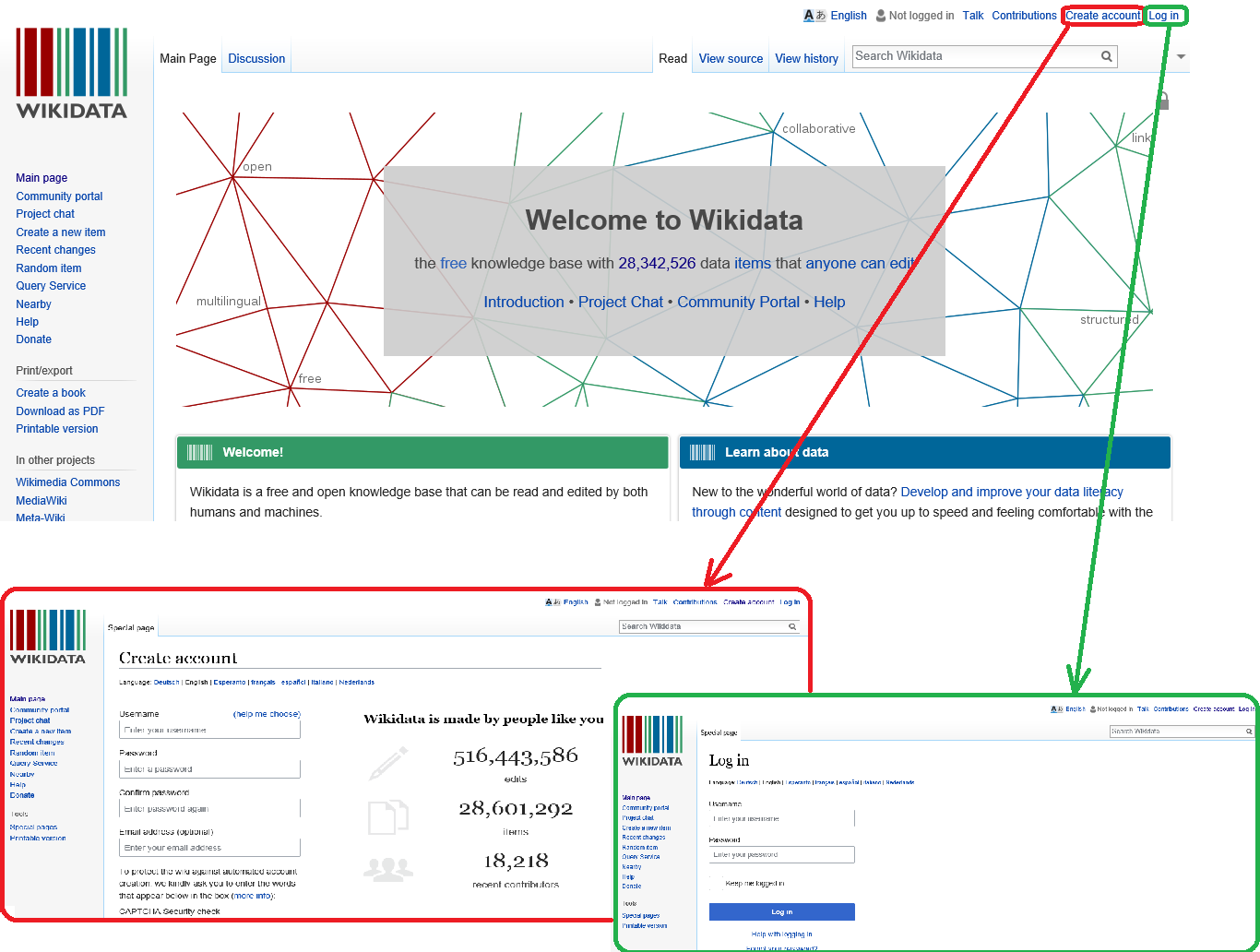
| Arabic dialect | ISO 639-3 code |
| --- | --- |
| Tunisian | AEB-ARAB |
| Algerian | ARQ |
| Moroccan | ARY |
| Egyptian | ARZ |
| North Levantine | APC |
| South Levantine | AJP |
| Cypriot | ACY |
| Libyan | AYL |
| Sudanese | APD |
| Saharan | AAO |
| Hassaniya | MEY |

As Arabic dialects do not have any officially recognized regulatory institution (excepting Derja Association that was created to regulate Tunisian) 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].

## 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.



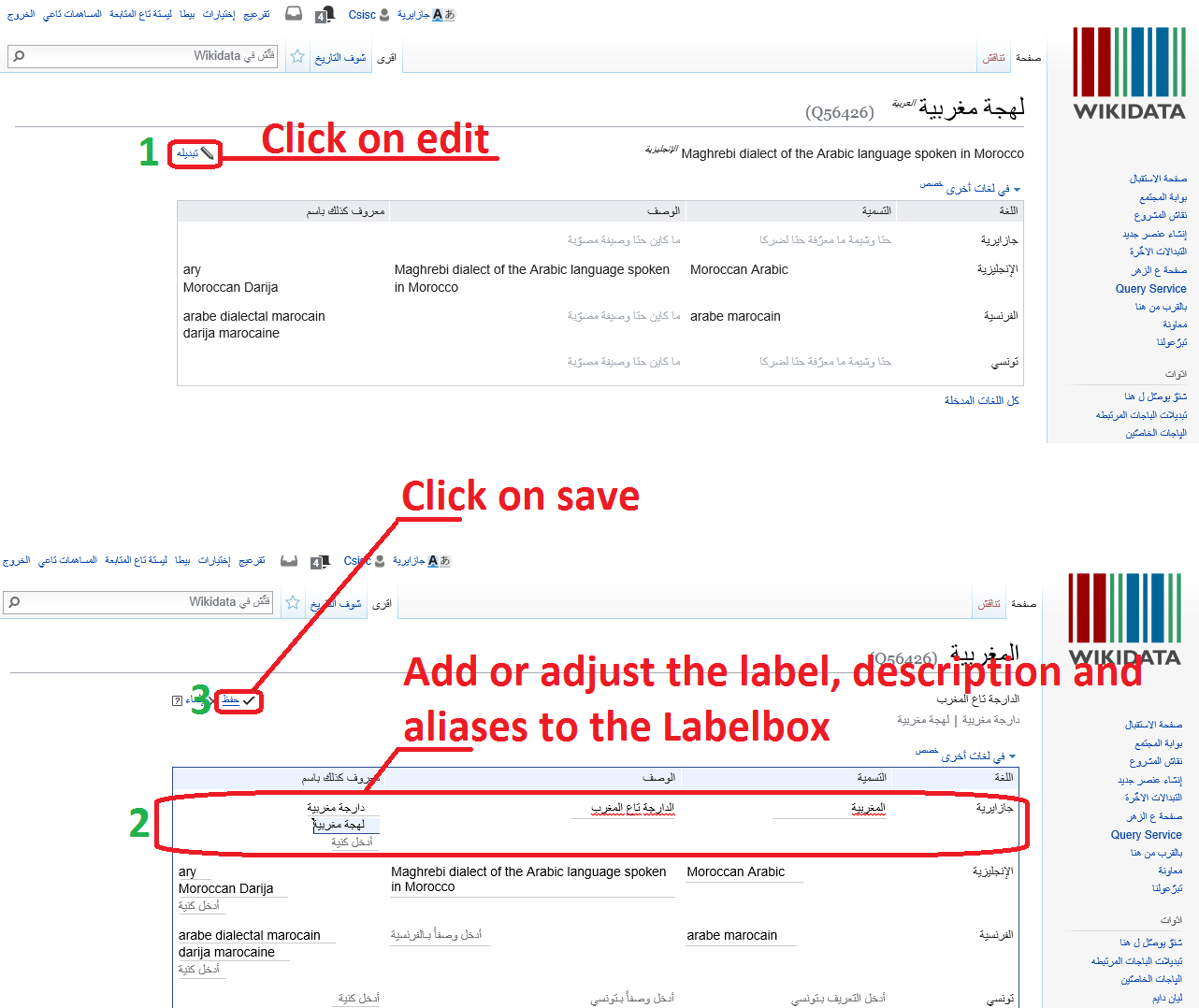
1. How to create a Wikimedia account [shown here in red] and how to log in to Wikidata [shown here in green] (Sources: <https://www.wikidata.org/wiki/Wikidata:Main_Page>, <https://www.wikidata.org/w/index.php?title=Special:CreateAccount>, and <https://www.wikidata.org/w/index.php?title=Special:UserLogin>,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

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



1. 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)

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.

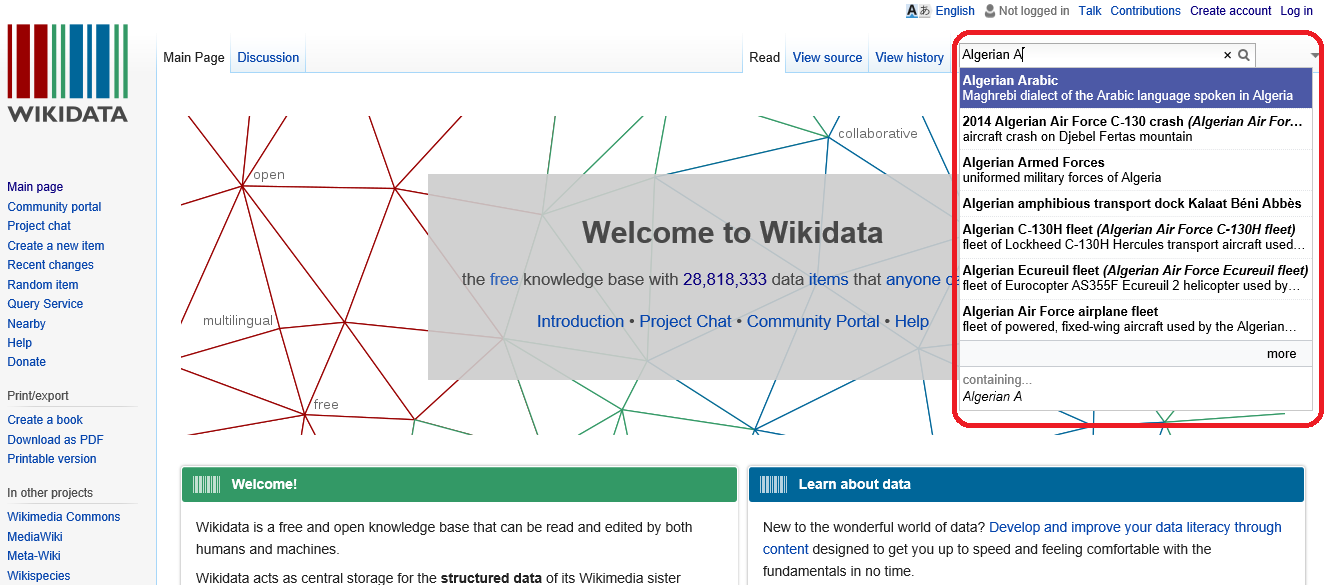


1. How to add or adjust the label in the chosen Arabic dialect through the page (Source: <https://www.wikidata.org/wiki/Q56426>, Credits: Houcemeddine Turki - CC BY-SA 3.0 License)

## 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.

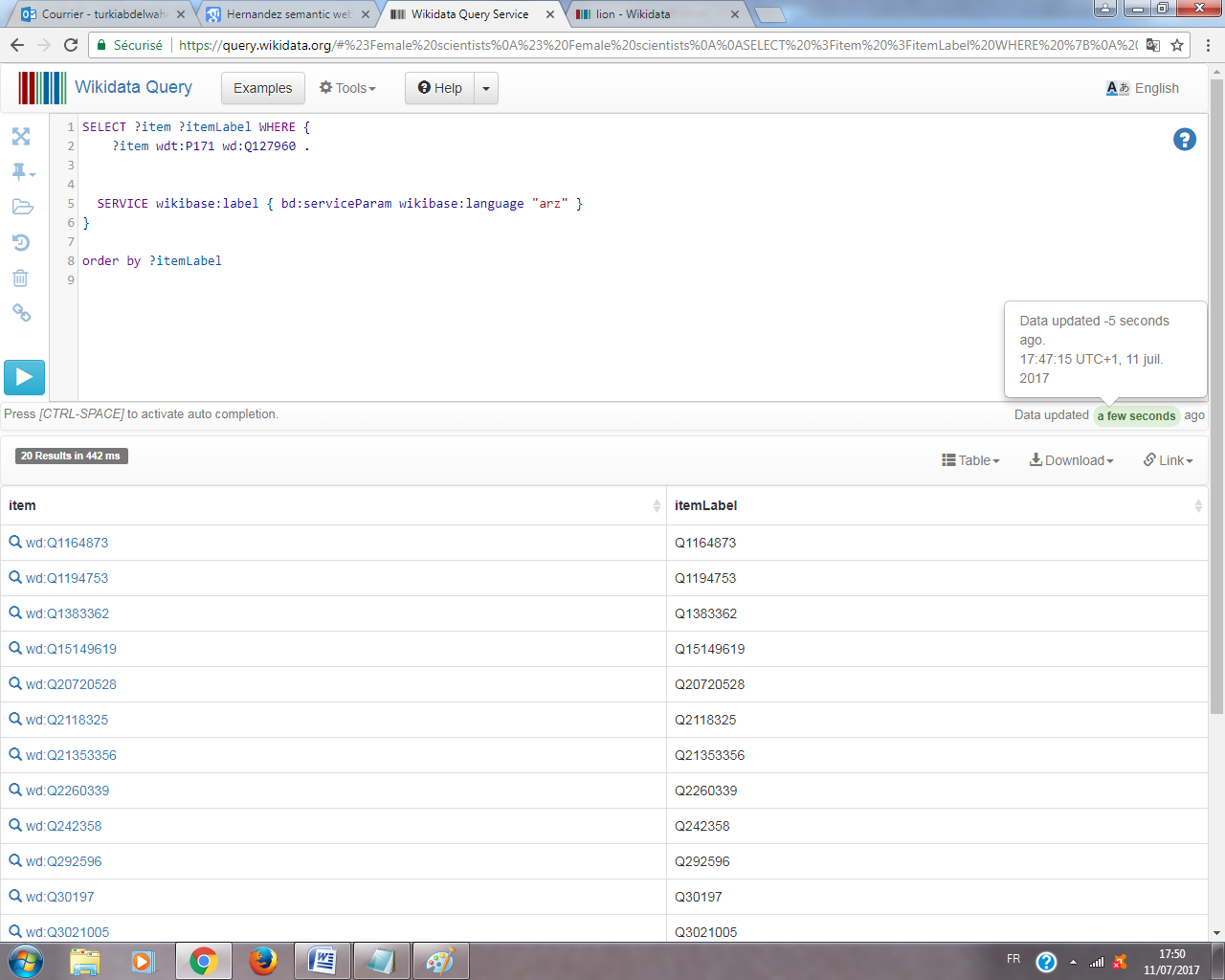


1. How to search a Wikidata item or property by typing its label in a chosen language in the search box of the main page of Wikidata (Source: <https://www.wikidata.org/wiki/Wikidata:Main_Page>, Credits: Houcemeddine Turki - CC BY-SA 3.0 License)

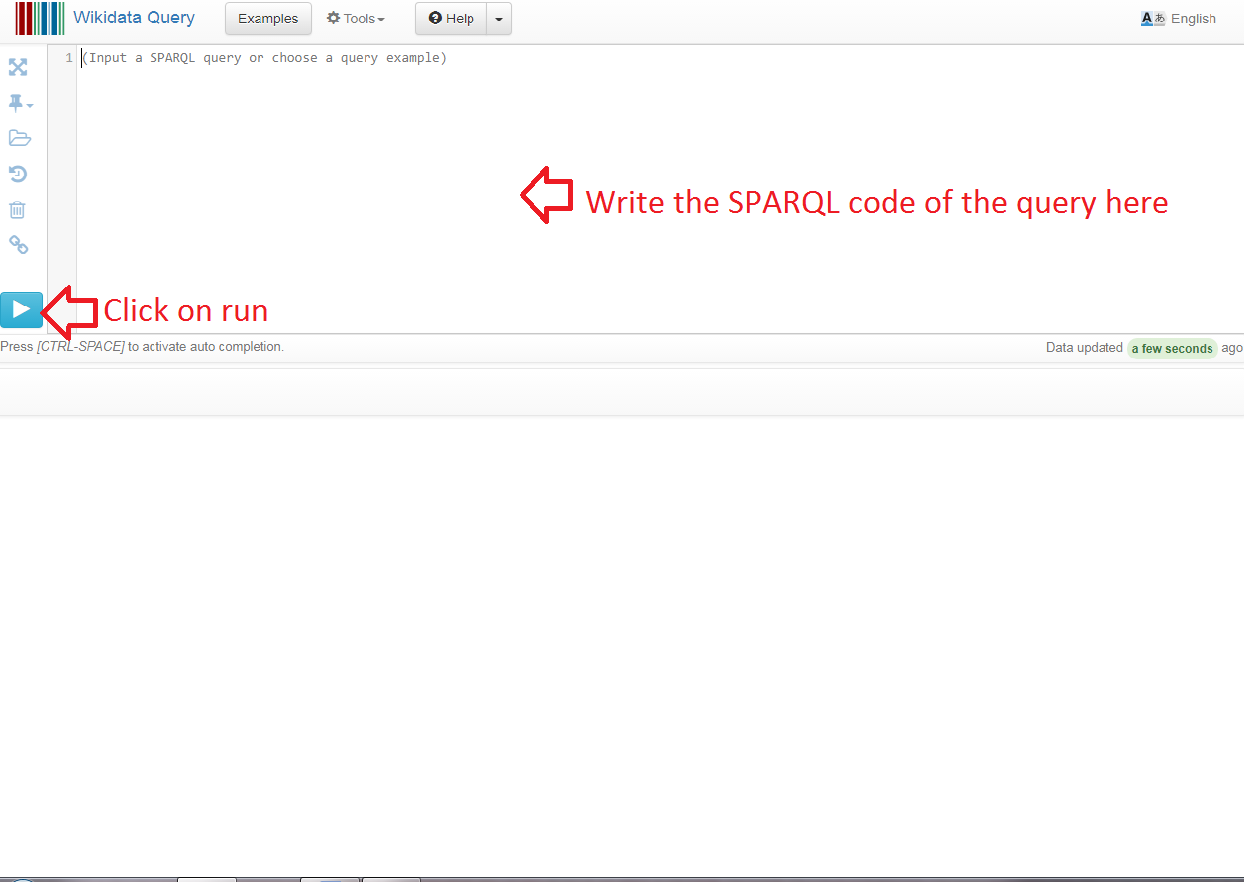
Consequently, you will access the page of the searched Wikidata item or property. You can find the name of the Wikidata item or property in the target language or Arabic dialect in the Labelbox of that page as shown in Fig. 3.

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).



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



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

# 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 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.

## 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.

Consequently, Wikidata can be used to find the false friends between Arabic dialects like بندق meaning “hazelnut” in Modern Standard Arabic and “pine nut” in Tunisian.

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].

## 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].

(English) 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) كامي لاكور هو عوّام فرانساوي

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

* **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 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].

Wikidata

Camille Lacourt (Q714898)

Place of Birth (P19): Narbonne

↓

(Label in Tunisian for Camille Lacourt) كامي لاكور

(Label in Tunisian for Narbonne) ناربون

↓

Most relevant search result containing the two labels:

كامي لاكور إتولد في ناربون

↓

Result: The expression used to give the place of birth of a given person: إتولد في

1. An example of how Wikidata is used by morphological analyzers to find a structure used in the analyzed language for a specific function

* **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].
* **Identififcation 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.

Message (in Tunisian): علي مركى زوز بونتوات و ربح الماتش

↓

Closely related terms : ماتش (match), بونتو (goal), مركى (to score), ربح (to win)

↓

When analyzing the semantic network of Wikidata, we found that all these terms are indirectly related to the term: sport

↓

Result: The general topic of the message is sport

1. An example of how Wikidata is used by identify the general topic of a text written in an Arabic dialect

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].

# Conclusion And Future Directions

In this research paper, we propose Wikidata (a free, collaborative, and multilingual knowledge database available in <https://www.wikidata.org>) as 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 betweeen semantic items as data in Wikidata is organized in RDF format.

Further than its common and well known direct functions as a multi-lingual multi-dialectal dictionary for Arabic dialects and just like other similar projects such as the Arabic dialect Treebank and the Arabic dialect WordNet, Wikidata can be also useful for other indirect functions like Machine Translation, POS Tagging, and the identification of the general topics of texts for sentiment analysis or semantic classification of texts.

That is why we believe that Wikidata is an outstanding contribution to Computational Linguistics that can be used as a database for other projects (apps, software, online interfaces...) for the NLP of Arabic dialects.

We think that future directions for this work will be the use of the data provided in Wikidata in tools like the ones for Machine Translation between Arabic dialects and the use of 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.

# 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 from Linguistic Data Consortium, University of Pennsylvania for his helpful comments and discussions.

# References

|  |  |
| --- | --- |
| [1] | R. Bassiouney, Arabic sociolinguistics, Edinburgh University Press, 2009. |
| [2] | J. Owens, The Oxford handbook of Arabic linguistics, Oxford University Press, 2013. |
| [3] | D. Cohen, "Les deux parlers arabes de Tunis," *Études de linguistique sémitique et arabe,* pp. 150-171, 1970. |
| [4] | M. T. Feghali, Syntaxe des parlers arabes actuels du Liban, Geuthner, 1928. |
| [5] | D. Cohen and M. el-Chennafi, Le dialecte arabe Hassaniya de Mauritanie, C. Klincksieck, 1963. |
| [6] | J. Grand'Henry, Les parlers arabes de la région du Mzāb:(Sahara algérien), Brill Archive, 1976. |
| [7] | F. A. Nureddeen, "Cross cultural pragmatics: Apology strategies in Sudanese Arabic," *Journal of pragmatics,* vol. 40, no. 2, pp. 279-306, 2008. |
| [8] | M. Farghal, "The pragmatics of’inšāllah in Jordanian Arabic," *Multilingua-Journal of Cross-Cultural and Interlanguage Communication,* vol. 14, no. 3, pp. 253-270, 1995. |
| [9] | S. Ghazali, R. Hamdi and M. Barkat, "Speech rhythm variation in Arabic dialects," in *Speech Prosody 2002, International Conference*, 2002. |
| [10] | J. Kaye, "Government in phonology. The case of Moroccan Arabic," *The linguistic review,* vol. 6, no. 2, pp. 131-160, 1987. |
| [11] | J. Heath, "Moroccan Arabic phonology," *Phonologies of Asia and Africa (including the Caucasus),* vol. 1, pp. 205-217, 1997. |
| [12] | M. Maamouri, The Georgetown Dictionary of Iraqi Arabic: Arabic-English, English-Arabic, Georgetown University Press, 2013. |
| [13] | M. Maamouri, The Georgetown Dictionary of Moroccan Arabic: Arabic-English, English Arabic, Georgetown University Press, 2017. |
| [14] | E. S. Badawi and M. Hinds, A Dictionary of Egyptian Arabic: Arabic-English, Librairie du Liban, 1986. |
| [15] | R. Ben Abdelkader, A. Ayed and A. Naouar, Peace Corps English-Tunisian Arabic Dictionary, Washington, DC: Peace Corps, 1977. |
| [16] | M. Diab, M. Al-Badrashiny, M. Aminian, M. Attia, P. Dasigi, H. Elfardy, R. Eskander, N. Habash, A. Hawwari and W. Salloum, "Tharwa: A Large Scale Dialectal Arabic-Standard Arabic-English Lexicon," in *LREC*, 2014. |
| [17] | 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. |
| [18] | D. Vrandecic, "The rise of Wikidata," *IEEE Intelligent Systems,* vol. 28, no. 4, pp. 90-95, 2013. |
| [19] | F. Erxleben, M. Günther, M. Krötzsch, J. Mendez and D. Vrandecic, "Introducing Wikidata to the Linked Data Web," in *Semantic Web Conference (1)*, 2014. |
| [20] | D. Vrandečić and M. Krötzsch, "Wikidata: a free collaborative knowledgebase," *Communications of the ACM,* vol. 57, no. 10, pp. 78-85, 2014. |
| [21] | V. Cavalli-Sforza, H. Saddiki, 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 (aiccsa), 2013 acs international conference on*, 2013. |
| [22] | R. Bouchlaghem, A. Elkhlifi and R. Faiz, "Tunisian dialect Wordnet creation and enrichment using web resources and other Wordnets," in *Proc. EMNLP Workshop ANLP*, 2014. |
| [23] | M. Maamouri, A. Bies, T. Buckwalter, M. Diab, N. Habash, O. Rambow and D. Tabessi, "Developing and using a pilot dialectal Arabic treebank," in *Proceedings of the Fifth International Conference on Language Resources and Evaluation, LREC’06*, 2006. |
| [24] | M. Maamouri, A. Bies, S. Kulick, M. Ciul, N. Habash and R. Eskander, "Developing an Egyptian Arabic Treebank: Impact of Dialectal Morphology on Annotation and Tool Development," in *LREC*, 2014. |
| [25] | E. Simon, I. Z. Benyeda, P. Koczka and Z. Ludányi, "Automatic creation of bilingual dictionaries for Finno-Ugric languages," in *Septentrio Conference Series*, 2015. |
| [26] | G. F. Simons and C. D. Fennig, Ethnologue: Languages of the world, SIL International, 2017. |
| [27] | N. Habash, M. T. Diab and O. Rambow, "Conventional Orthography for Dialectal Arabic," in *LREC*, 2012. |
| [28] | I. Zribi, R. Boujelbane, A. Masmoudi, M. Ellouze, L. H. Belguith and N. Habash, "A Conventional Orthography for Tunisian Arabic," in *LREC*, 2014. |
| [29] | H. Saadane and N. Habash, "A conventional orthography for Algerian Arabic," in *ANLP Workshop 2015*, 2015. |
| [30] | N. Y. Habash, M. T. Diab and O. C. Rambow, Conventional Orthography for Dialectal Arabic (CODA): Principles and Guidelines--Egyptian Arabic-Version 0.7-March 2012, Columbia University Academic Commons, 2014. |
| [31] | M. Maamouri, A. Bies, S. Kulick, D. Tabessi and S. Krouna, Egyptian Arabic Treebank DF Parts 1-8 V2. 0-LDC catalog numbers LDC2012E93, Linguistic Data Consortium, 2012. |
| [32] | H. Turki, E. Adel, T. Daouda and N. Regragui, "A Conventional Orthography for Maghrebi Arabic," in *Proceedings of the International Conference on Language Resources and Evaluation (LREC), Portoroz, Slovenia*, 2016. |
| [33] | M. Jarrar, N. Habash, D. Akra and N. Zalmout, "Building a corpus for palestinian arabic: a preliminary study," in *Proceedings of the EMNLP 2014 Workshop on Arabic Natural Language Processing (ANLP)*, Doha, Qatar, 2014. |
| [34] | S. Khalifa, N. Habash, D. Abdulrahim and S. Hassan, "A Large Scale Corpus of Gulf Arabic," in *Language Resources and Evaluation Conference*, 2016. |
| [35] | M. Maamouri, D. Graff, H. Jin, C. Cieri and T. Buckwalter, "Dialectal Arabic Orthography-based Transcription", in *EARS RT-04 Workshop*, 2004. |
| [36] | B. Zawaydeh, D. Stallard and J. Makhoul, "Babylon Transcription Guidelines", 2003. [Online]. Available: http://ldc.upenn.edu/Catalog/docs/LDC2005S08/BBNBabylontranscription-guidelines.pdf. [Accessed: 11- Jul- 2017]. |
| [37] | D. Hernández, A. Hogan, C. Riveros, C. Rojas and E. Zerega, "Querying Wikidata: comparing SPARQL, relational and graph databases", in *International Semantic Web Conference*, 2016, pp. 88-103. |
| [38] | F. Daouas, A. Ben Abdelaziz, T. Ajmi and A. Mtiraoui, "Les attitudes professionnelles humaines des médecins de la région sanitaire de Sousse (Tunisie)", *Santé publique*, vol. 14, no. 2, pp. 135-145, 2002. |
| [39] | R. Zbib, E. Malchiodi, J. Devlin, D. Stallard, S. Matsoukas, R. Schwartz and C. Callison-Burch, "Machine translation of Arabic dialects", in *Proceedings of the 2012 conference of the north american chapter of the association for computational linguistics: Human language technologies*, 2012, pp. 49-59. |
| [40] | E. Rabinovich, S. Mirkin, R. Patel, L. Specia and S. Wintner, "Personalized Machine Translation: Preserving Original Author Traits", in *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, 2017, pp. 1074–1084. |
| [41] | J. Hirschberg and C. Manning, "Advances in natural language processing", *Science*, vol. 349, no. 6245, pp. 261-266, 2015. |
| [42] | M. Klang and P. Nugues, "Langforia: Language Pipelines for Annotating Large Collections of Documents", in *COLING (Demos)*, 2016, pp. 74-78. |
| [43] | A. Thompson, "Using part-of-speech tags to identify presence of location information in social media messages", in *Southern California Conferences for Undergraduate Research*, 2016. |
| [44] | M. Abdul-Mageed, M. Diab and S. Kübler, "SAMAR: Subjectivity and sentiment analysis for Arabic social media", *Computer Speech & Language*, vol. 28, no. 1, pp. 20-37, 2014. |