Knowledge Graphs: Semantics, Machine Learning, and Languages
M. Acosta et al. (Eds.)

doi:10.3233/SSW230004

QALD-9-ES: A Spanish Dataset for Question Answering Systems

Javier SORUCO [a,1], Diego COLLARANA [a,b,2], Andreas BOTH [c,d,3] and Ricardo USBECK [e,4]

[a] *Universidad Privada Boliviana, Cochabamba, Bolivia*
[b] *Fraunhofer FIT, Sankt Agustin, Germany*
[c] *Leipzig University of Applied Sciences, Leipzig, Germany*
[d] *DATEV eG, Nuremberg, Germany*
[e] *Universität Hamburg, Hamburg, Germany*

Abstract. Knowledge Graph Question Answering (KGQA) systems enable access to semantic information for any user who can compose a question in natural language. KGQA systems are now a core component of many industrial applications, including chatbots and conversational search applications. Although distinct worldwide cultures speak different languages, the number of languages covered by KGQA systems and its resources is mainly limited to English. To implement KGQA systems worldwide, we need to expand the current KGQA resources to languages other than English. Taking into account the recent popularity that Large-Scale Language Models are receiving, we believe that providing quality resources is key to the development of future pipelines. One of these resources is the datasets used to train and test KGQA systems. Among the few multilingual KGQA datasets available, only one covers Spanish, i.e., QALD-9. We reviewed the Spanish translations in the QALD-9 dataset and confirmed several issues that may affect the KGQA system's quality. Taking this into account, we created new Spanish translations for this dataset and reviewed them manually with the help of native speakers. This dataset provides newly created, high-quality translations for QALD-9; we call this extension QALD-9-ES. We merged these translations into the QALD-9-plus dataset, which provides trustworthy native translations for QALD-9 in nine languages, intending to create one complete source of high-quality translations. We compared the new translations with the QALD-9 original ones using language-agnostic quantitative text analysis measures and found improvements in the results of the new translations. Finally, we compared both translations using the GERBIL QA benchmark framework using a KGQA system that supports Spanish. Although the question-answering scores only improved slightly, we believe that improving the quality of the existing translations will result in better KGQA systems and therefore increase the applicability of KGQA w.r.t. the Spanish language domain.

Keywords. Knowledge Graphs, Question Answering, Dataset

[1] Javier Soruco mail: javiersorucol1@upb.edu
[2] Diego Collarana mail: diego.collarana.vargas@iais.fraunhofer.de
[3] Andreas Both mail: andreas.both@datev.de
[4] Ricardo Usbeck mail: ricardo.usbeck@uni-hamburg.de

1. Introduction

The main goal of question-answering systems (QA systems) is to provide access to knowledge graphs via natural language, saving users from learning a specific graph query language to retrieve information from KGs. To achieve this goal, researchers have created different components and tools to mature the KGQA systems. These tools include benchmarking datasets to measure the quality of KGQA systems and datasets such as LC-QuAD [1] or QALD-9 [2] to train different KGQA components. Although natural language is the perfect medium for a comfortable experience for the end user, it also restricts who can take advantage of these systems. Recent developments in KGQA systems have heightened the need for multilingual tools and components. Ideally, KGQA systems should be available in various languages, making them accessible to diverse cultures. However, most KGQA research has focused mainly on English, leaving aside a significant number of languages, some of which are spoken by millions of people, e.g., Spanish, which is spoken by approximately 427 million people and is the world's second-most spoken native language[5].

QALD-9 is one of the few multilingual datasets that facilitate the development of KGQA systems in 11 languages. At the moment of writing this paper, QALD-9 is the only multilingual dataset available that provides Spanish translations. Unfortunately, most of the translations in QALD-9 are grammatically incorrect and unnatural[6]. Spanish is not the exception; after our analysis, we have found that the quality of Spanish translations of QALD-9 that have existed so far is relatively low. These issues go from poorly written translations to cases where the meaning of the original question is lost.

QALD-9-plus [3] has addressed this problem by improving the quality of these translations with the help of native speakers. QALD-9-plus adds translations in languages that were not included in the original benchmark, creating a dataset with high-quality translations available for nine different languages (en, de, fr, ru, uk, lt, be, ba, hy) and two knowledge graphs: Wikidata[7] and DBpedia[8].

In order to develop reliable KGQA systems for Spanish, the availability of high-quality resources that allow for training and testing of the systems becomes essential. We hope that improving the quality of a Spanish dataset will result in improved KGQA system performance for the given language. In this work, we aim to extend QALD-9-plus to include one additional language – Spanish. To achieve this goal, we manually created new translations with the help of native Spanish speakers. We also evaluated the results using language-agnostic quantitative text analysis measures and the tool GERBIL QA [4] to compare the results of the original translations and the new translations; we named the new translations "QALD-9-ES".

We address the problem of providing KGQA tools in multiple languages and propose QALD-9-ES, a KGQA dataset based on QALD-9 that contains accurate Spanish translations. We integrate QALD-9-ES with QALD-9-plus, complementing this multilingual dataset containing accurate natural translations with Spanish. Extending the scope of trustful translations for a dataset is essential for creating multilingual systems that serve a diverse population. In summary, the contributions of this work are as follows:

[5] cf. https://www.ethnologue.com/statistics/summary-language-size-19
[6] https://github.com/ag-sc/QALD/issues/22
[7] https://www.wikidata.org/
[8] https://www.dbpedia.org/

- A three-fold process to analyze KGQA datasets, i.e., Qualitative Analysis, Translations Review, and Quantitative Analysis. We apply this process to analyze the quality of the Spanish language in the QALD-9 dataset.
- A publicly available KGQA dataset with accurate Spanish translations. We integrate our work with QALD-9-plus to increase its adoption.
- An evaluation of our QALD-9-ES dataset against KGQA systems using the GERBIL QA framework and showing improvements in most of the metrics compared to its predecessor.

This article is organized into the following sections: (2) previous work, (3) dataset development and description, (4) baseline evaluation, and (5) conclusions.

2. Previous Work

For the elaboration of this work, we reviewed KGQA datasets and related tools to develop and compare KGQA datasets. Table 1 summarizes each dataset's information showing the lack of accurate Spanish translations.

Table 1. For existing KGQA datasets, we show the number of unique questions and available languages. QALD-9 is the only dataset available for the Spanish language, but it suffers from quality issues.

Dataset	Available for	No. of questions	Available languages
QALD-9	DBpedia	558	en, it, de, ru, fr, pt, hi_IN,fa, ro, es, nl
QALD-9-plus[9]	DBpedia Wikidata	558	en, lt, de, ru, fr, uk, be, ba, hy
rewordQALD9	DBpedia	551	en, it
LC-QUAD	DBpedia	5000	en
LC-QUAD 2.0	DBpedia Wikidata	30000	en

2.1. KGQA Datasets

2.1.1. QALD-9, the 9th Challenge on Question Answering over Linked Data (QALD-9):

QALD is a challenge with eleven years of history with the objective of providing up-to-date benchmarks for assessing and comparing state-of-the-art KGQA systems[10]. QALD-9 [2] is the 9th edition of the QALD challenge. This dataset provides 408 training questions and 150 test questions for DBpedia, available in 11 different languages, making QALD-9 one of the few multilingual KGQA benchmarks available and the only one we are aware of that counts with Spanish translations. In the 9th version of QALD, the questions were compiled and curated from previous versions and are accompanied by manually specified SPARQL queries and answers. The community reported multiple issues with the translations in the QALD-9 dataset; they were reported to be incorrect and of poor quality for different languages [3]. After reviewing the Spanish translations, we

[9]The number of questions differs depending on the language
[10]https://www.nliwod.org/challenge

found that Spanish was not the exception. We detail the issues found with the Spanish translations in Section 3.1. The QALD-9 dataset works with the QALD-JSON data format, which allows multiple languages, and it is used as a communication format with systems like GERBIL QA [4].

2.1.2. *QALD-9-plus, a Multilingual Dataset for Question Answering over DBpedia and Wikidata Translated by Native Speakers:*

QALD-9-plus [3] is an initiative to fight the lack of multilingual KGQA benchmarks and the translation issues of QALD-9. This dataset provides an extended version of QALD-9 with 4,930 new question translations for different languages. The translations were done via crowdsourcing. Each crowd worker was assigned a subset of QALD-9 questions to translate into their mother tongue, resulting in at least two translations per question. Crowd workers were later given two translations and the original question, and they had to decide whether the first or second translation was correct or whether both or no translations were correct. The QALD-9-plus dataset also includes a version of the dataset for Wikidata that was generated manually by three experienced computer scientists with the help of semi-automatic scripts to speed up the process. The result is an extended version of the QALD-9 dataset available for nine languages (English included) with high-quality translations, adjusted to work with both DBpedia and Wikidata knowledge graphs.

2.1.3. *RewordQALD9, a Bilingual Benchmark with Alternative Rewordings of QALD Questions:*

rewordQALD9 [5] is an extended version of the QALD-9 dataset that brings forward high-quality Italian translations with multiple reformulations for the same question. Rewordings are available for both Italian and English; therefore, testing systems' robustness is available for both languages. The translations were manually curated by native speakers, including reformulations for both English and Italian. The resulting dataset consists of 551 questions in both English and Italian. In addition, multiple question reformulations are included, i.e., 1546 for English and 1707 for Italian.

2.1.4. *LC-QuAD, a Corpus for Complex Question Answering over Knowledge Graphs:*

LC-QuAD [1] is the solution to the necessity of large datasets composed of various question templates and their logical forms for QA systems. LC-QuAD is generated based on an entity seed list and a predicate whitelist to obtain subgraphs from DBpedia. Then the graphs are used to generate a SPARQL from a template, which generates a natural language question from Normalized Natural Question Templates. Finally, the questions are manually reviewed and corrected. LC-QuAD is composed of 5000 questions, and the SPARQL queries are required to answer the questions on DBpedia. It was one of the most extensive datasets available for KGQA at the time of release. This dataset is only available in English; therefore, it is not multilingual.

2.1.5. *LC-QuAD 2.0, a Large Dataset for Complex Question Answering over Wikidata and DBpedia:*

LC-QuAD 2.0 [6] is the second version of LC-QUAD, providing 30,000 questions with their paraphrases and corresponding SPARQL queries. LC-QuAD 2.0 is compatible with Wikidata and DBpedia 2018 knowledge graphs. To generate the dataset, the authors gen-

erated SPARQL queries based on templates, which were then transformed into template questions. By using crowdsourcing, the template questions were verbalized into natural-language questions. This dataset, like its predecessor, is only available in English.

2.2. Tools Related to KGQA Systems and Benchmarks

2.2.1. Benchmarking Question Answering Systems:

GERBIL QA [4] is an online benchmarking platform for question-answering systems (derived from the GERBIL tool for evaluating entity recognition approaches, cf. [7]). GERBIL QA follows the FAIR principles to provide a quality evaluation of QA systems. This platform allows users to benchmark their systems with relevant datasets such as QALD and LC-QuAD. GERBIL QA also allows its users to upload their datasets. The platform is connected to relevant KGQA solutions so that the users can compare their systems with the relevant systems available; these systems can work with any private dataset uploaded by the users only if this dataset uses the QALD-JSON format. GERBIL QA offers seven metrics for benchmarking QA systems and supports online and file-based systems.

2.2.2. Question Answering Benchmark Curators

QUANT [8] is a framework for creating or curating QA benchmarks, generating smart edit suggestions for questions-query pairs and their metadata, and providing predefined quality checks for queries. QUANT reduces the curation effort for QA benchmarks by up to 91%. QUANT is a suitable tool during the KGQA dataset development process, e.g., QALD-9 used QUANT.

3. QALD-9-ES Dataset Development and Description

3.1. Qualitative Analysis of QALD-9

The QALD-9 dataset comprises 558 questions, of which 408 correspond to the training dataset and 150 to the testing dataset. Each question contains a list of question objects for every available language. A question object is composed of the question string and the question keywords. The question string is the question expressed in a given language (e.g., Spanish), and the question's keywords are key elements of the question that the KGQA system can use as support to answer questions for the given language. These keywords are usually related (but not restricted) to proper entity names, verbs, nouns, and adjectives.

As we mentioned before, QALD-9 already has Spanish translations; after an explorative review, we concluded that the quality of the Spanish translations is doubtful. Thus, we decided to review all the original translations, classifying each translation into seven cases.

Cases 1 to 6 correspond to error cases, which can be split into cases with errors in the question string (cases 1, 2, 5, and 6), or cases with mistakes in the question keywords (cases 3, 4), while case 7 implies a correct translation.

Each question can be assigned to multiple cases, with the only exception being case 7. If a sentence follows into Case 7, it implies a correct translation; therefore, it cannot be

assigned to questions with translation errors in the question string, but it can be assigned to questions with mistakes in the question keywords.

1. The first case involves what we call "minor translation mistakes," which we define as translation errors that do not alter the question's original meaning. Some common errors, in this case, are the use of the incorrect genders (for example, using the word "hermoso" for a female noun or subject), missing words that do not contribute to the meaning of the question, the absence of opening and closing question marks ('¿?'), missing letters, the inappropriate use of plural forms, the lack of capital letters in proper names, or using the wrong tense. An example of these issues can be found within the question "Who developed Skype?", which got translated to "Quien desarrollado Skype?"; The verb "desarrollado" is in its past participle form. To be a correct translation, the auxiliary verb "ha" should accompany the main verb resulting in the translation "¿Quién ha desarrollado Skype?". Another correct alternative is to modify the main verb by changing it to the simple past tense, resulting in "¿Quién desarrolló Skype?".
2. The second case is about what we call "major translation errors", the main characteristic of these errors is that they result in the loss of meaningful elements of the question, such as verbs, proper names, and other meaningful words. One example of this case is found in the question "In which U.S. state is Area 51 located?" which got translated to "En cual Nosotros estado es Zona 51 ¿situado?". If we reverse this translation, we would get "In which us state is Zone 51 located?" As you can see, the entity "U.S." was mistranslated to the word "us" and the entity "Area 51" was transformed to "Zone 51". In both cases, the entities in question were lost, meaning that a KGQA system would not be able to work properly with this translation. Another example of this issue can be found in the question "Where did Abraham Lincoln die?" that gets translated to "Dónde hizo Abrahán Lincoln el?". In this translation, the verb "die" was completely lost, resulting in a question that just does not make sense. In the original QALD-9 translations, we found some questions that lacked their corresponding Spanish translation. Those questions were classified into this case. This case can be triggered by mistakes both in the question and in the question's keywords.
3. The third case relates to questions where the question's keywords require modifications due to some mistakes in the question's original translation that propagate into the question's keywords. We can find this case in the question "What is the last work of Dan Brown?", which got translated to "Qué es el último trabajo de Y ¿Marrón?", In this case, the entity Dan Brown was mistranslated to "Y ¿Marrón?", resulting in error propagation into the question keywords "último trabajo, Y marrón". The correct keywords for this question are "último trabajo, Dan Brown".
4. The fourth case is similar to the third case; the difference is that the errors found in the question's keywords are not related to errors in the question's translation. For example, words are correctly written in the question's translation but wrongly written in the question's keywords. In the question "Which monarchs were married to a German?", we find that the Spanish keywords for the question are "monarcha, casado, alemán". The word "monarcha" is wrongly translated; the correct translation "monarca". This mistranslation was only found in the question's keywords; the translation lacks this mistake.

5. The fifth case is about Spanish accentuation. In Spanish, the character "´ " is known as "orthographic accent". This accent is used on some words' vowels and can modify the word's meaning. One example is the words "mas" and "más". "mas" is an equivalent of the word "but", while "más" is a quantity adverb. Some QALD-9 translations lack orthographic accents, so the real meaning of some words is lost. This case can be triggered by mistakes both in the question and in the question keywords.
6. The sixth case comprises questions with correct translations that were reformulated to be more natural. Modifications in the question's keywords do not trigger this case. The question "How deep is Lake Placid?" was translated to "Cómo de hondo es el Lago Placid?", which is technically a correct translation, but we consider that the literal translation "Cómo de hondo" is not the best way to express "How deep"; therefore, we modified the translation to "¿Qué tan profundo es el Lago Placid?".
7. Finally, the seventh case relates to questions that do not require modification. Modifications to the question's keywords do not affect this case. The question "Which presidents were born in 1945?" was translated to "¿Qué presidentes nacieron en 1945?". We consider this translation correct and natural; therefore, we classified it under Case 7.

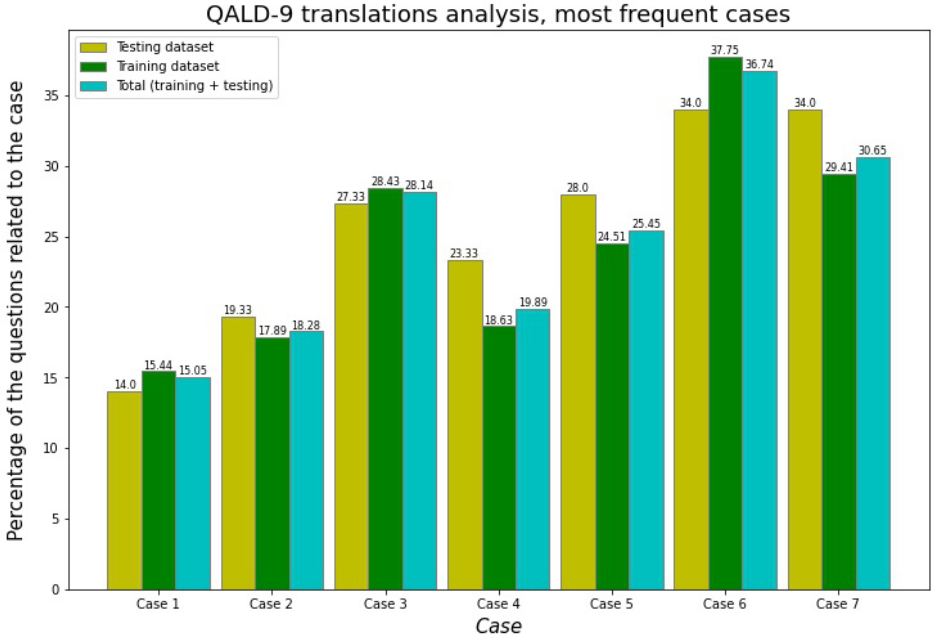

Figure 1. Plot showing the percentage of occurrence of each case in the QALD-9 dataset. Case 1 is related to questions with minor translation issues, case 2 to major translation issues resulting in the loss of the original meaning of the question, case 3 relates to errors on the question keywords propagated from the question translation, case 4 to mistakes in the question keywords that are not related to errors in the question translation, case 5 is about questions with accentuation errors, case 6 contains questions that were modified to be more natural, and finally, case 7 are questions with correct translations.

Figure 1 shows the percentage of occurrence of each case, 15.05% of the question presented minor translation mistakes (Case 1), 18.28% of the questions had major translation issues (Case 2), 28.14% of the question's keywords had mistakes that resulted from errors in the question's translation (Case 3). 19.89% of the question keywords had errors that were not related to errors in the question's translation (Case 4), 25.45% of the question presented accentuation errors in the question string or the question keywords (Case 5), 36.74% of the questions were modified due to their lack of naturalness (Case 6), and 30.65% of the questions were considered correct translations. We also found 23 questions lacking Spanish translation; these were classified as major translation issues.

3.2. Translation of Questions

The translation process consisted of two agents, "the translator" and "the reviewer", interacting within two stages: the translation stage and the review stage.

The translator is a native Spanish speaker with the main objective of generating new translations while matching the old translations to one of the seven cases described before. This agent is also required to have a high level of understanding of SPARQL and the QALD-JSON format in order to be able to check the question information if required.

The reviewer is a native Spanish speaker with the main objective of reviewing the new translation and providing feedback to the translator to improve the quality of the new translations.

In the translation stage, the translator reviews each question by checking the original English question, its original QALD-9 translation, the English question keywords, and the QALD-9 translation keywords. The translator annotates the cases that describe the question translation (see above) and generates a translation for the English question. The translation is kept if the question's translation is correct (case 7). The translator will always look for the Spanish version of proper names in order to keep the translation as natural as possible (cases like "Iraq", which is spelled "Irak" in Spanish); this also implies not translating proper English names if the Spanish-speaking community knows the entity by the English name (like the TV show "Friends"). The translator's goal is always to generate correct translations (avoid grammatical and accentuation mistakes) that fit in the natural Spanish dialect.

In the review stage, the reviewer checks the original English question, the translation generated by the translator, and the keywords from the new translation, looking for possible mistakes. The mistakes are annotated and sent to the translator. The translator reviews the observations, and if the translator agrees with the correction, the question is modified; corrections that raise additional concerns about the translation are annotated and discussed with the reviewer. The reviewer explains why the corrections are necessary so the translator can make a choice; in some cases, the translator reviews the question's SPARQL query and answers to make an appropriate choice. This process was executed over the original QALD-9 dataset. Once the native translations were generated, they were merged into QALD_9_plus using the question's id, generating what we call QALD-9-ES[11].

[11] https://github.com/KGQA/QALD_9_plus

Table 2. Number of unique questions available for Wikidata and DBpedia in the QALD-9-plus extension QALD-9-ES.

	en	de	ru	uk	lt	be	ba	es	hy	fr
DBpedia train	408	543	1203	447	468	441	284	408	80	260
DBpedia test	150	176	348	176	186	155	117	150	20	26
Wikidata train	371	497	1095	407	426	403	260	371	71	251
Wikidata test	136	159	318	160	166	141	107	136	19	25

Table 3. Results of the linguistic evaluation of QALD-9 and QALD-9-ES Spanish translations performed by using LinguaF. The linguistic indicators used to compare the translations are average words per sentence, average word length, average syllable per word, lexical density, and type-token ratio.

	QALD-9 Train	QALD-9-ES Train	QALD-9 Test	QALD-9-ES Test
Average words per sentence	7.044706	7.504808	6.767442	7.434211
Average word length	5.034402	5.097053	5.058419	5.114159
Average Syllable per word	1.418838	1.426650	1.416667	1.417699
Lexical density	78.423514	79.724536	76.804124	79.292035
Type Token Ratio	0.376420	0.356502	0.472509	0.465487

3.3. Dataset Statistics

The resulting dataset contains questions for DBpedia and Wikidata Knowledge Graphs. In Table 2, we show that the dataset contains 408 train questions for DBpedia, 371 train questions for Wikidata, 150 test questions for DBpedia, and 136 test questions for Wikidata in Spanish. The rest of the languages are preserved as in QALD-9-plus [3].

3.4. Quantitative Analysis between QALD-9 vs QALD-9-ES

Inspired by the work of the QALD-9-plus team, we have used language-agnostic quantitative text analysis measures to observe the differences between the Spanish translations of QALD-9 and QALD-9-ES. To achieve this, we used the library "LinguaF"[12].

The results can be observed in Table 3. QALD-9-ES' translations have more words; each word is longer and has more syllables. QALD-9-ES also improves the dataset's lexical density, meaning that the dataset has more meaningful words (e.g., nouns, verbs, adjectives, and some adverbs). On the other hand, we have found that the new translations present a lower Type Token Ratio (TTR). That means there are fewer unique words in the new translations than in the old ones. After analyzing and comparing both translations, we hypothesized that this was because most questions that presented translation errors (like Case 1 and 3) tended to add incorrect and unrelated words. Questions related to Case 5 affect this measure too, as words that have errors related to accentuation in some questions are considered unique words by LinguaF.

[12] https://github.com/WSE-research/LinguaF

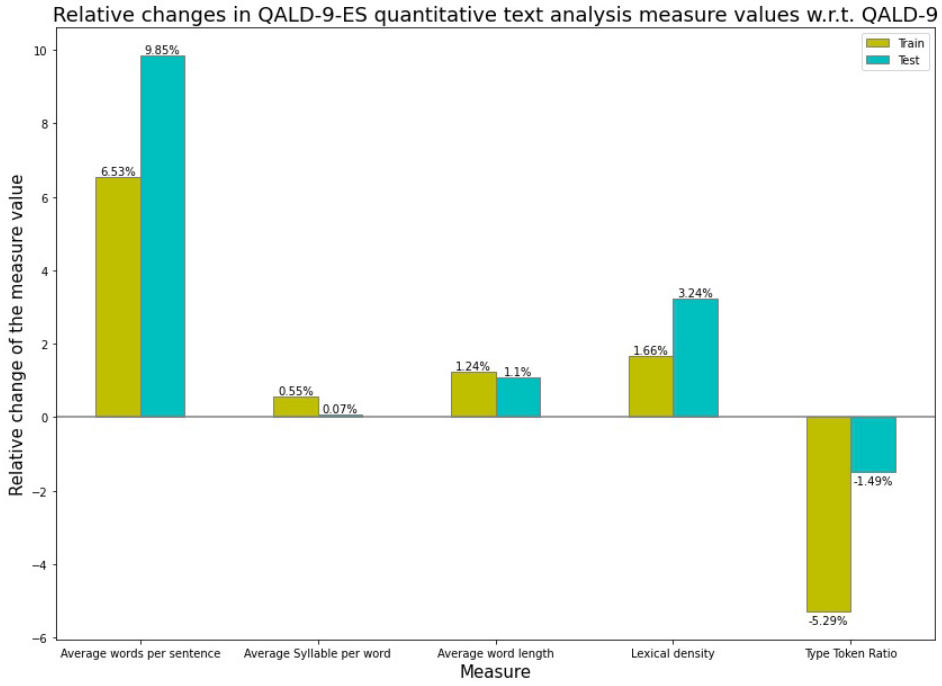

Figure 2. We use LinguaF to perform the linguistic evaluation, showing relative improvements in the Spanish translations comparing QALD-9 vs. QALD-9-ES. The linguistic indicators used are average words per sentence, average word length, average syllable per word, lexical density, and type-token ratio.

In Figure 2, we can see that the measures with the most improvement in QALD-9-ES are average words per sentence and lexical density, followed by average word length and average syllable per word, making TTR the only measure with a decrement in comparison to QALD-9.

4. Baseline Evaluation

With the premise that better translations result in better QA systems, we used the GERBIL QA system to evaluate the QALD-9-ES dataset by comparing it to QALD-9. At the time of writing this paper, QAnswer [9] is the only working annotator available on GERBIL QA that supports Spanish QA. QAnswer is available in two versions, one that works with DBpedia KG and the other with Wikidata KG. While working with DBpedia, both QALD-9-ES and QALD-9 resulted in system errors; therefore, we only compared the datasets using QAnswer over Wikidata. QALD-9 does not include a version for Wikidata; hence, we created a version by replacing the QALD-9-plus translations with the original QALD-9 translations in the Wikidata set.

When running the experiments on GERBIL QA, we received not only the QA results but the results of three sub-experiments that can measure the quality of a QA system; Resource to Knowledge Base (C2KB), Properties to Knowledge Base (P2KB) and Relation to Knowledge Base (RE2KB). The results for QALD-9-ES are shown in Table 5, and for QALD-9 in Table 4. For each sub-experiment, we present the F1 score metric.

Table 4. Results of the GERBIL QA evaluation performed on QALD-9 Spanish, using the QAnswer annotator over Wikidata. GERBIL performs four sub-experiments: Question Answering (QA), identification of relevant resources (C2KB), identification of relevant properties (P2KB), and matching of expected triples (RE2KB).

Dataset	Sub-experiment	Micro F1	Macro F1
QALD-9 test	QA	0.1077	0.1522
	C2KB	0.3798	0.3408
	P2KB	0.4183	0.4069
	RE2KB	0.1141	0.1691
QALD-9 train	QA	0.1588	0.2486
	C2KB	0.3682	0.3538
	P2KB	0.4082	0.4106
	RE2KB	0.1703	0.2276

Table 5. Results of the GERBIL QA evaluation performed on QALD-9-ES using the QAnswer annotator over Wikidata. GERBIL performed four sub-experiments: Question Answering (QA), identification of relevant resources (C2KB), identification of relevant properties (P2KB), and matching of expected triples (RE2KB). We see improvements in most of the scores when using QALD-9-ES.

Dataset	Sub-experiment	Micro F1	Macro F1
QALD-9-ES test	QA	0.1142	0.1680
	C2KB	0.3893	0.3593
	P2KB	0.4093	0.3951
	RE2KB	0.1236	0.1863
QALD-9-ES train	QA	0.1775	0.2639
	C2KB	0.3623	0.3471
	P2KB	0.3804	0.3887
	RE2KB	0.1874	0.2419

The F1 score is the harmonic mean between the system's precision and recall presented in Eq. (1). This metric is applied over one class; in this case, we have several classes; therefore, the micro and macro average aggregation methods are applied. Macro F1 is the unweighted mean of the F1 scores obtained per class presented in Eq. (2). The Micro F1 method calculates the F1 score using the normal F1 equation but using the total number of True Positive (TP), False Positive (FP), and False Negative (FN) values instead of the values of a single class and it is presented in Eq. (3). The main difference between the micro and macro metrics is that Micro F1 gives equal importance to each observation; consequently, some classes will significantly impact the results for imbalanced datasets. On the other hand, Macro F1 gives equal importance to the class F1 score, allowing it to return objective results even on imbalanced datasets.

$$F1 = \frac{TP}{TP + \frac{1}{2} * (FP + FN)} \tag{1}$$

$$Macro\ F1 = \frac{\sum F1\ scores}{Number\ of\ classes} \tag{2}$$

$$Micro\ F1 = \frac{\sum TP}{\sum TP + \frac{1}{2} * (\sum FP + \sum FN)} \tag{3}$$

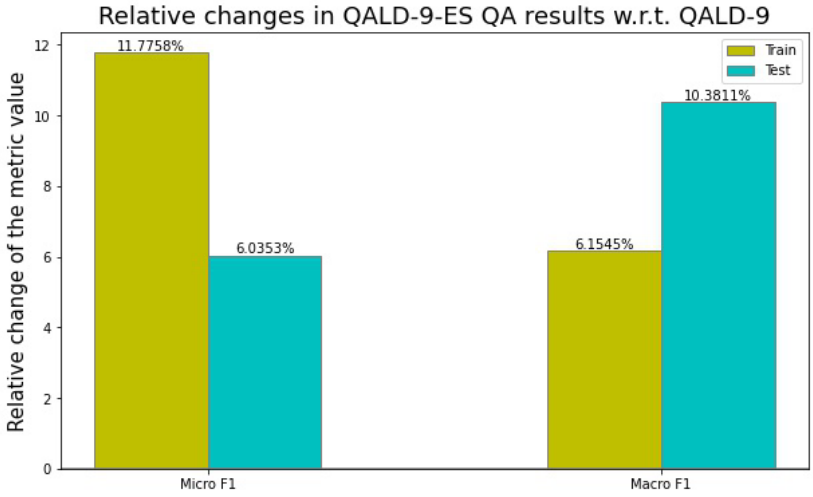

Figure 3. Relative changes in the question answering (QA) results of QAnswer over Wikidata using the QALD-9-ES Spanish translations compared to the same annotator using the QALD-9 Spanish translations.

The results of the QA experiment show an increment in both Micro and Macro F1 (Figure 3) when using QALD-9-ES. The Micro F1 measure value increased more in the training set than in the testing set, and Macro F1 shows a more significant improvement in the testing set than in the training set.

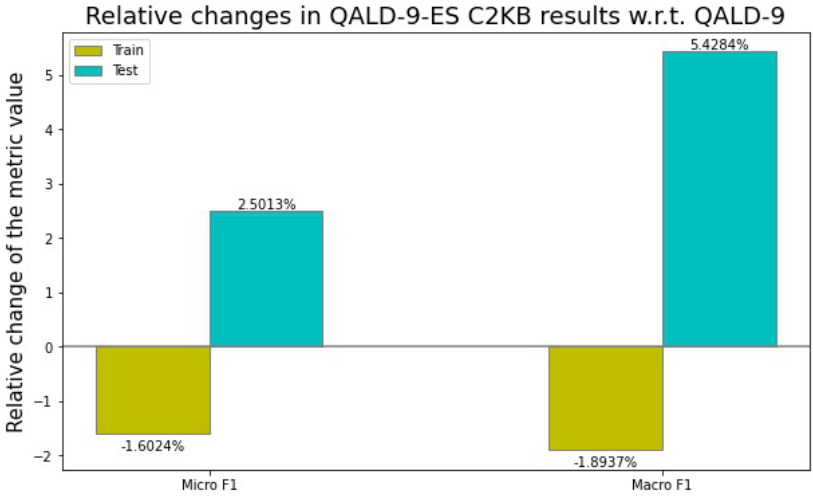

Figure 4. Relative changes in the GERBIL sub-experiment C2KB results of QAnswer over Wikidata using the QALD-9-ES Spanish translations compared to the same annotator using the QALD-9 Spanish translations.

The C2KB sub-experiment qualifies the capability of the system to identify all the relevant resources for the given question. Figure 4 shows a relative increment in Micro and Macro F1 scores for the testing set, but the scores decrement for the training set.

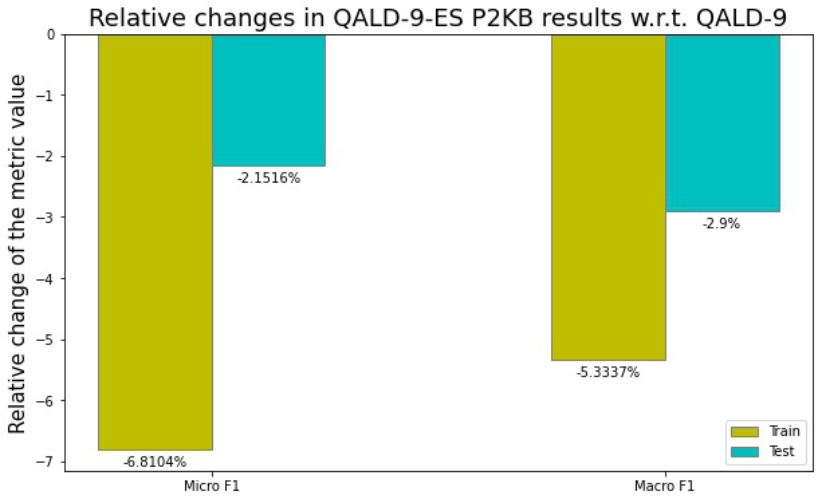

Figure 5. Relative changes in the GERBIL sub-experiment P2KB results of QAnswer over Wikidata using the QALD-9-ES Spanish translations compared to the same annotator using the QALD-9 Spanish translations.

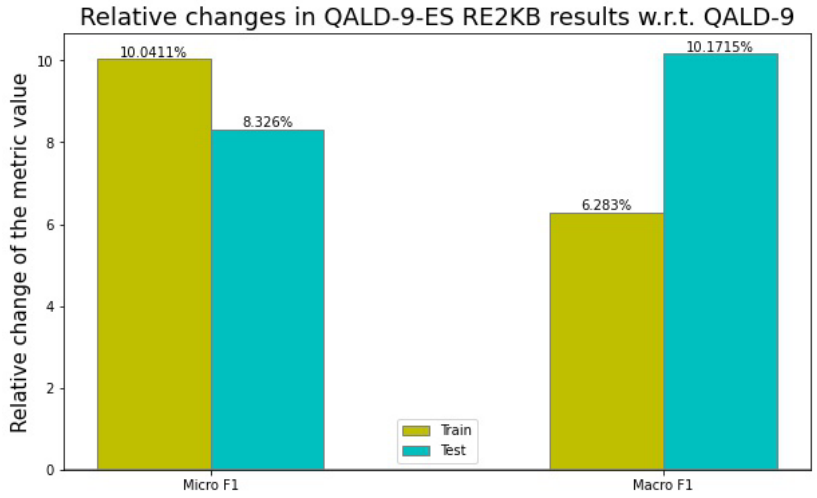

Figure 6. Relative changes in the GERBIL sub-experiment RE2KB results of QAnswer over Wikidata using the QALD-9-ES Spanish translations compared to the same annotator using the QALD-9 Spanish translations.

The P2KB sub-experiment qualifies the system's capability to identify all the relevant properties for the given question. In this case, we see a decrement for Micro and Macro F1 for both the testing and training datasets (Figure 5). Finally, the RE2KB sub-experiment compares the expected triples in the question's expected SPARQL against the triples in the SPARQL returned by the QA system. QALD-9-ES shows an improvement in this sub-experiment compared to QALD-9 in both the testing and training sets for Micro and Macro F1 (cf. Figure 6).

5. Conclusions

KGQA systems provide access to knowledge graph information through natural language. However, the number of unique natural languages is not comparable to the number of languages covered by existing KGQA systems. This paper addresses the problem of providing multilingual tools to develop KGQA systems to increase the number of languages these systems cover. We focus on Spanish, a language spoken by more than 450 million people worldwide.

Following a three-fold approach, we performed a qualitative analysis of the Spanish translations presented in QALD-9 and found that only 30.65% of the Spanish translations in QALD-9 properly represent the English questions. Then, manually generate new translations for the questions that presented translation issues and review them manually with the help of native speakers. We integrate QALD-9-ES with QALD-9-plus, a QALD-9-based dataset made exclusively with native translations, so there is a complete source of high-quality translations for QALD-9 that can be used for the development of new datasets and KGQA systems.

We compared the QALD-9-ES Spanish translations with the original translation included in QALD-9 using language-agnostic quantitative text analysis measures to confirm that the new translations use more words, each word is longer, and there are more meaningful words in each translation. The only downside is that there are fewer unique words. After some review, we hypothesize that this result is because the original QALD-9 translations often included unrelated words in the translations and several accentuation mistakes that are taken as unique words.

Using the GERBIL QA framework, we also evaluate both datasets, i.e., QALD-9-ES vs QALD-9. We use the QAnswer annotator for Wikidata (the only annotator working in Spanish KGQA at the time of evaluation). The experiment results show minor improvements in the QA results. The system also showed slightly better results in identifying relevant resources (C2KB) for the testing set and slightly worse results for the training set. We also found that the annotator has lower performance in identifying the relevant properties for a given question (P2KB) with the new translations, but the annotator has better results in matching the expected triples for each question (RE2KB). These mixed results show the impact of the quality of the dataset on the KGQA system and components.

Finally, we demonstrated how the QALD-9-ES dataset is useful for the development of Spanish KGQA pipelines. We expect that this new dataset will especially benefit KGQA systems that use Large-Scale Language Models in their pipeline. The resulting dataset of this work was merged into QALD-9-plus, a fully native translated dataset, this resource can also be used to compare native translations against translations generated using Machine Translation.

Acknowledgement: This work is supported by the Research Partnership Grant RPG2106 funded by the Swiss Leading House for Latin America, and by grants for the DFG project NFDI4DataScience project (DFG project no. 460234259) and by the Federal Ministry for Economics and Climate Action in the project CoyPu (project number 01MK21007G).

References

[1] Trivedi P, Maheshwari G, Dubey M, Lehmann J. LC-QuAD: A Corpus for Complex Question Answering over Knowledge Graphs. In: The Semantic Web - ISWC 2017 - 16th International Semantic Web

Conference, Vienna, Austria, October 21-25, 2017, Proceedings, Part II. vol. 10588 of Lecture Notes in Computer Science. Springer; 2017. p. 210-8.

[2] Ngomo N. 9th challenge on question answering over linked data (QALD-9). vol. 7; 2018. p. 58-64.

[3] Perevalov A, Diefenbach D, Usbeck R, Both A. QALD-9-plus: A Multilingual Dataset for Question Answering over DBpedia and Wikidata Translated by Native Speakers. In: 2022 IEEE 16th International Conference on Semantic Computing (ICSC); 2022. p. 229-34.

[4] Usbeck R, Röder M, Hoffmann M, Conrads F, Huthmann J, Ngomo AN, et al. Benchmarking question answering systems. Semantic Web. 2019;10(2):293-304.

[5] Sanguinetti M, Atzori M, Puddu N. rewordQALD9: A Bilingual Benchmark with Alternative Rewordings of QALD Questions. In: Proceedings of Poster and Demo Track and Workshop Track of the 18th International Conference on Semantic Systems co-located with 18th International Conference on Semantic Systems (SEMANTiCS 2022), Vienna, Austria, September 13th to 15th, 2022. vol. 3235 of CEUR Workshop Proceedings. CEUR-WS.org; 2022. .

[6] Dubey M, Banerjee D, Abdelkawi A, Lehmann J. LC-QuAD 2.0: A Large Dataset for Complex Question Answering over Wikidata and DBpedia. In: The Semantic Web - ISWC 2019 - 18th International Semantic Web Conference, Auckland, New Zealand, October 26-30, 2019, Proceedings, Part II. vol. 11779 of Lecture Notes in Computer Science. Springer; 2019. p. 69-78.

[7] Usbeck R, Röder M, Ngonga Ngomo AC, Baron C, Both A, Brümmer M, et al. GERBIL: General Entity Annotator Benchmarking Framework. In: Proceedings of the 24th International Conference on World Wide Web. WWW '15. Republic and Canton of Geneva, CHE: International World Wide Web Conferences Steering Committee; 2015. p. 11331143. Available from: https://doi.org/10.1145/2736277.2741626.

[8] Gusmita RH, Jalota R, Vollmers D, Reineke J, Ngomo AN, Usbeck R. QUANT - Question Answering Benchmark Curator. In: Semantic Systems. The Power of AI and Knowledge Graphs - 15th International Conference, SEMANTiCS 2019, Karlsruhe, Germany, September 9-12, 2019, Proceedings. vol. 11702 of Lecture Notes in Computer Science. Springer; 2019. p. 343-58.

[9] Diefenbach D, Giménez-García J, Both A, Singh K, Maret P. QAnswer KG: Designing a Portable Question Answering System over RDF Data. In: Harth A, Kirrane S, Ngonga Ngomo AC, Paulheim H, Rula A, Gentile AL, et al., editors. The Semantic Web. Cham: Springer International Publishing; 2020. p. 429-45.