RESEARCH ARTICLE OPEN ACCESS

Construction of ACL Ontology and ACL lexicon for developing semantic analyzer for an Arabic Controlled Language

Hoyam Salah Elfahal Elebaed [1], Karim BOUZOUBAA [2]

[1] College of Post-graduate Studies, Sudan University of Science and Technology,Khartoum, Sudan

[2] Computer Science Department, Mohammadia School of Engineers, Mohammed V University in Rabat - Morocco

ABSTRACT

The Arabic Controlled Languages (ACL) is a subset of Arabic natural languages with specific rules and vocabulary. The aim of such a controlled language is to develop Arabic NLP applications with reduced complexity and ambiguity. However, computers cannot understand the meaning of ACL sentences. In this paper, and relying on a previously build ACL corpus, we introduce an ACL lexicon and an ACL ontology for the development of an ACL semantic analyzer. We also developed an ACL vocabulary checker to verify the words of any ACL sentence according to the ACL lexicon.

Keywords: - Arabic Controlled Language, ACL, Natural language processing, ACL lexicon, ACL ontology and ACL vocabulary checker.

I. INTRODUCTION

Arabic digital content has grown at breakneck speed over the past decade. The number of Arabic-speaking Internet users has increased by 9,348.0% in the last twenty years[1] . This increase is due to the interest and importance of the Arabic language being a Semitic language [1] and the fourth most spoken language in the world by more than 467 million people. This results in a growing need for more advanced NLP services for people and manufacturers. However, working on Arabic NLP is more challenging [2] due to its complexity and ambiguity and the lack of linguistics tools and resources [3].

The complexity of Arabic is recognized by considering both conservative, templatic and many accidental forms in a few morphological features [4].To alleviate this problem, we proposed in a previous work an Arabic-controlled language (ACL) [7] to make Arabic texts easier for computers to analyse. In the first step in the development of ACL, we defined the grammar rules and constructed a corpus of sentences with their corresponding grammar [5]. Consequently, users are able to check whether an Arabic sentence respects the ACL syntax. However, users are still unable to assess and benefit from semantic processing of ACL texts. Indeed, Schwitter et al. 2010 stated that an ideal machine-oriented CNL should have well-defined and precise semantics that is defined by an unambiguous mapping into a logic-based representation [6]. Therefore, we need to develop a semantic analyzer for ACL so that the computer can understand the meaning of ACL sentences. To this end and since a semantic analyzer needs specific resources, we first introduce the ACL lexicon and the ACL ontology. Also, we introduce an ACL vocabulary checker to verify the correctness of the vocabulary of sentences according to the ACL vocabulary. Such tool is necessary when developing applications relying on the ACL language such as machine translation. When the user types a sentence containing a term not present in the ACL vocabulary, the checker warns the user to provide an alternative word. The ACL Vocabulary Checker thus reduces training time and costs, and improves translation quality.

In Section 2, we highlight existing NLP infrastructures, resources, and tools to make the best choice for developing semantics for ACL. Section 3 shows how the ACL lexicon and ontology are constructed. Section 4 presents the software used to check the ACL vocabulary. Section 5 concludes this paper.

II. LITERATURE REVIEW

In this section, we conduct a literature review of NLP linguistics resources and NLP tools to make the best choice for developing an ACL ontology and the vocabulary checker.

2.1 ONTOLOGY

In general, the term ontology is used to refer to the study of the kinds and relations of things that exist in a given domain. Nasri et al. (2016) count 7 works that have shown the importance of ontologies in various fields such as 1) NLP 2) Information Retrieval (IR) 3) Information System 4) Semantic Web 5) Machine learning 6) Data mining and 7) Knowledge representation. For NLP, ontologies have been used to measure the semantic distance between concepts and to compute the semantic similarity between semantic graphs to provide a semantic representation of the meanings of key concepts and to support semantic reasoning [12]. To develop an ontology, there are 6 parts in the life cycle mentioned by Buitelaar et al. in 2005 such as 1) Creation 2) Population 3)

[1] https://www.internetworldstats.com/stats7.html

Validation, 4) Deployment 5) Maintenance, and 6) Evolution [13].

To build an ontology, there are different tools that can be used such as Amine's ontology/KB GUI[2], Protégé[3], Jena[4], KOAN[5], Sesame[6], NeOn Toolkit[7], SWOOP[8], Neologism[9], TopBraid Composer[10] , etc.

Majdi Beseiso et al. [14] evaluate many tools and found that all the evaluated systems do not support Arabic language processing or diacritics. However, there was no mention of the Amine ontology [11] that supports the Arabic language development.

In addition, for Western languages such as English and Spanish, there are many open-source ontologies that belong to either the specific or the open domain category: such as OpenCyc [15], Know-ItAll [16], HowNet [11], SNOMED[12], GeneOntology [13], etc. As for the Arabic language, a different picture emerges. To our knowledge, the first work covering both lexical terms and semantic representations of concepts is by Abouenour et al [17]. Their ontology is structured around a concept hierarchy, lexical information and semantic frames (called situations) related to these concepts. The former of their ontology are extracted from the Arabic Wordnet (AWN) [8] while the latter are represented based on the transformation of the Arabic Verbnet (AVN) [9][10]. The level of Amine's Arabic ontology consists of a) Situations, related to verb situation concepts such as Relation Linguistic, Object, State, and Situation. b) action_root, contains the elementary actions used by AVN as predicate values (desire, transfer_info, perceive, etc. c) Manner, an empty branch to contain the manners. d) The فعل and اسم branches contain all verbs and nouns extracted from AVN and AWN, respectively.

Figure 1: The top level of the Amine Arabic ontology.

2.3 NLP INFRASTRUCTURES

n the last decade, a number of multilingual infrastructures [18] have been developed, such as toolkits (a set of tools in a single box used for a specific purpose), platforms (consisting of several interoperable tools with a homogeneous structure, but not providing an API to extend their components), and frameworks (a multi-layered structure developed as a support and guide for building NLP programs and tools). Some of them are dedicated only to the Arabic language, such as.

1) PHARAS[14]: is a platform for parsing texts in modern standard Arabic. It includes several resources (dictionaries and lexicons) that are created semi-automatically. The syntactic subsystem covers verbal and nominal sentences.
2) AraNLP library[15] is a Java-based toolkit for Arabic NLP. It supports important preprocessing steps.
3) MADAMIRA is a Java toolkit for NLP processing of Arabic and its dialects presented by Pasha et al. in 2014 [19]. In addition, it provides services such as: Tokenization, Morphological Analysis, Disambiguation, Part-of-Speech Tagging, Lemmatization, Diacritization, Named Entity Recognition, Base Phrase Chunking.
4) AL-Khalil A set of applications named in general "Al-khalil" is developed by Natural Language Processing team at Mohammed First University[16]. Alkhali Suite offers a range of applications: Alkhalil Morphology analyzer, Alkhalil POS tagger, Alkhalil diacritizer ,Alkhalil stemmer, Alkhalil Lemmatizer, Almus'haf corpus , Tree Tagger. It also provides resources such as Al-Mus'haf corpus and the Nemlar one.
5) FARASA[17] Arabic is a Java text processing library developed by QCRI (Qatar Computing Research

[2] http://amine-platform.sourceforge.net/
[3] http://protege.stanford.edu
[4] https://en.wikipedia.org/wiki/Apache_Jena
[5] https://seco.cs.aalto.fi/tools/koan/
[6] https://www.w3.org/2001/sw/wiki/Sesame
[7] http://neon-toolkit.org
[8] http://www.mindswap.org/2004/SWOOP
[9] http://neologism.deri.ie
[10] https://www.topquadrant.com/products/topbraid-composer/
[11] www.keenage.com/html/e_index.html
[12] www.snomed.org
[13] www.geneontology.org

[14] https://www.loukam.net
[15] https://sites.google.com/site/mahajalthobaiti/resources
[16] http://oujda-nlp-team.net
[17] https://farasa.qcri.org/

Institute). It provides services such as 1) segmentation/tokenization module 2) POS tagger Arabic text diacritizer 3) dependency parser 4) named entities recognizer 5) spell checker. In addition, it provides an open-source online demo. It is also mentioned as a fast and accurate text processing toolkit for Arabic text.

6) SAFAR [20] is a Java-based natural language processing framework. It is a monolingual framework developed in accordance with software engineering requirements and targeted at the Arabic language, specifically modern standard Arabic and the Moroccan dialect. As shown in Figure 2, SAFAR has several layers that provide services that can be directly used by other layers according to the relationships modelled with arrows in Figure 2. 2) Tools (includes a set of technical services and pre-processing tools, as well as machine and deep learning utilities) 3) Resources (provides services for maintaining, consulting, and managing Arabic language resources such as corpora, dictionaries, and ontologies) 5) Application (includes high-level applications such as sentiment analysis or question/answer systems) 6) Client Applications (interacts with all other layers to serve clients through web applications, web services, etc.)

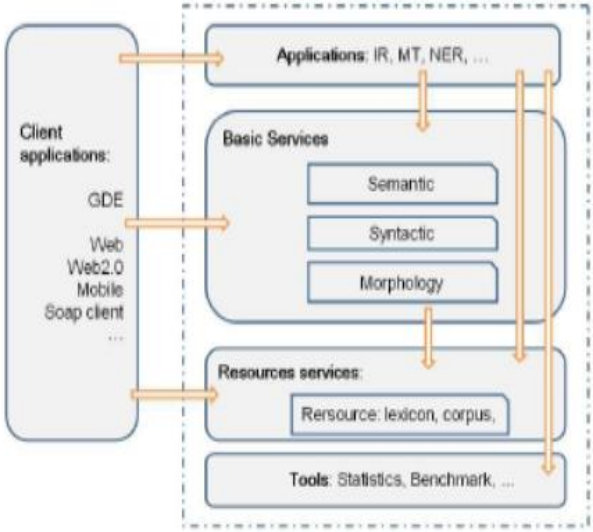

Figure 2: SAFAR framework general architecture.

III. RESOURCES FOR DEVELOPING SEMANTIC ANALYZER OF ACL

The following section propose linguistic resources such as the ACL lexicon and the ACL ontology to be used for the development of the ACL semantic analyzer.

3.1 BUILDING THE ARABIC CONTROLLED LANGUAGE LEXICON

The ACL lexicon is the vocabulary of the Arabic CNL. Its main purpose is to provide information about words for ACL sentences such as nouns, verbs, and adjectives. To build the ACL lexicon, we use the ACL corpus presented in a previous work [5]. Let's briefly recall that the ACL corpus was collected from textbooks and websites for teaching kids. It contains 551 sentences, all of which conform to the ACL grammar rules, such as "هذه ممسحة كبيرة."/This is a big mop). As shown in Figure 3, we extracted all the vocabulary from the ACL by pre-processing all the sentences, including tokenization, stop word removal, lemmatization, and then we removed redundant ones. Table 1 shows an example.

.

TABLE 1: OVERVIEW OF THE ACL SENTENCE "يشتد الحرّ في الصيف."

PREPROCESSING

Sentence	Processing type	An ACL Sentence Processing
"يشتد الحرّ في الصيف." \ It gets hot in the summer."	Tokenization	"يشتد الحرّ في الصيف"
	Stop word removal	"يشتد الحرّ الصيف "
	Lemmatization	"اِشْتَدّ حَرٌّ صَيْف"

Then, we obtain the ACL vocabulary containing nouns[18] (a word that refers to a person, place, object, event, substance, idea or feeling such as 'محمد', 'مسجد', 'كتاب', and 'سوق'.), verbs[19] (a word that refers to an action, state, or experience such as 'ذهب', 'جلس', and 'صاد'), and adjectives[20] (a word that describes a noun or pronoun such as 'كبير', 'طويل', and 'أحمر'.) (see Table 2) that we will later use to develop the

[18] https://dictionary.cambridge.org/dictionary/english-arabic/noun
[19] https://dictionary.cambridge.org/dictionary/english-arabic/verb
[20] https://dictionary.cambridge.org/dictionary/english-arabic/adjective?q=adjectives

ACL ontology.

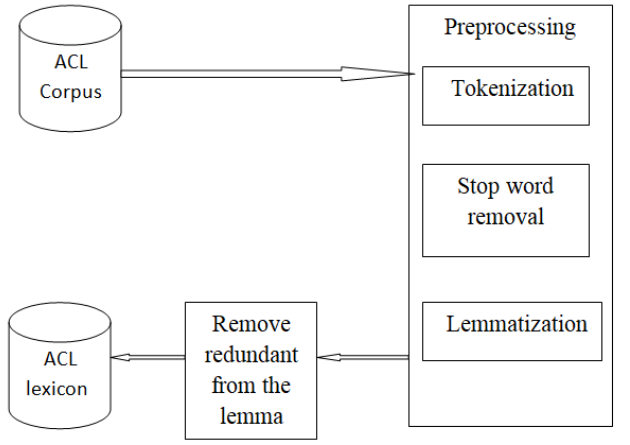

Figure 3. Overview of the pre-processing model of ACL corpus

TABLE 2: OVERVIEW OF THE ACL LEXICON WITH THE NUMBER OF VOCABULARY FOR EACH TYPE

Type	Number of Vocabulary for each type
Noun	213
Verb	99
Adjective	42
Total	354

3.2 BUILDING THE ARABIC CONTROLLED LANGUAGE ONTOLOGY

To provide the semantic information for ACL sentences, we need an ACL ontology. To do so, we enriched the Arabic ontology developed by Abouenour in 2014 [17] with the vocabulary from the ACL vocabulary. Considering that each lemma in the ACL vocabulary is a concept, we searched for these concepts in the Amine Arabic ontology to check whether they exist or not. For example, the ACL vocabulary " مسجد" "mosque" is not found in the Amine Arabic ontology (Figure 4). Therefore, we assume that this concept does not exist in the Arabic ontology. The concept "مسجد" is then manually added in the ontology.

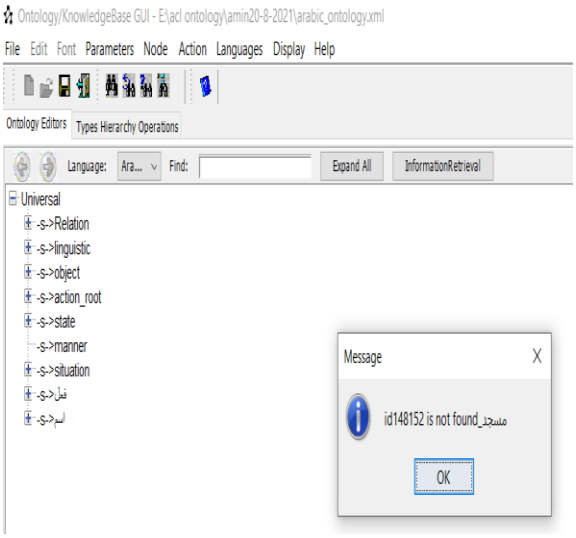

Figure 4. The concept " مسجد " not found in the Amine Arabic ontology

Table 3 shows statistics of the ACL vocabulary that exists the Amine Arabic ontology.

TABLE 3: STATISTICS OF THE ACL VOCABULARY THAT EXISTS IN THE AMINE ARABIC ONTOLOGY

Type	Exist	does not exist
Noun	73	140
Verb	75	24
Adjective	2	40
Total	150	204

Then we search for super type for ACL vocabulary that don't exist in Amine Arabic ontology. For example, the ACL vocabulary "سوق, مسجد, متحف," "market, museum, mosque" don't exist in the Arabic ontology but are all places "مكان". Therefore, we consider "مكان/place" as a super type of these concepts.

Finally, we constructed an ACL ontology with 42616 concepts by adding 57.62% of ACL concepts (vocabularies) that do not exist in the Arabic ontology according to the corresponding supertype.

Figure 5 shows the result of searching for the concept " مسجد " after inserting it into the ACL ontology.

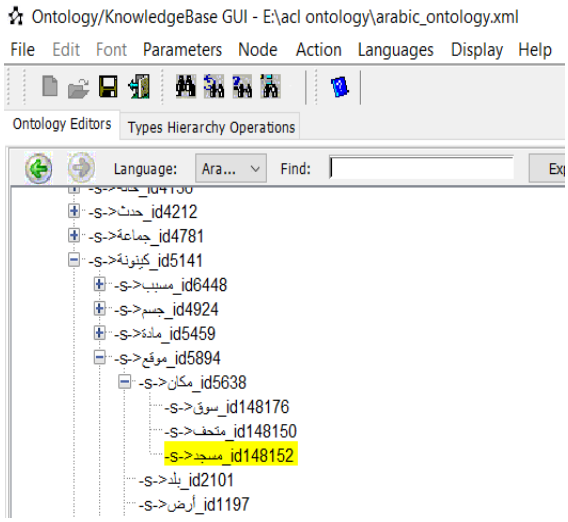

Figure 5. The concept "مسجد " found after add to Amine Arabic ontology

IV. THE VOCABULARY CHECKER

Although controlled languages can be implemented simply as a set of written guidelines for authors, the consistent quality of controlled language texts is maximized when the author uses a controlled language checker to write texts that are checked to see if they conform to the controlled language definition. Therefore, many controlled language checkers have been introduced to ensure that texts conform to the desired vocabulary and grammatical constraints, such as ClearCheck for CTE [21], Teruko et al. [22] for KANT Controlled English, and Hoard et al. [23] for other examples of checkers for CE. Thus, in this section, we present an ACL vocabulary checker.

The main idea behind the ACL vocabulary checker is to check whether the vocabulary of an input sentence matches the ACL vocabulary or not. For this purpose, the vocabulary checker performs pre-processing of the input sentence, such as normalization, tokenization, removal of stop words, determining the lemmas of the words, and then compares them with all ACL lemmas. Figure 6 provides an overview of the architecture of the ACL vocabulary checker. First, the user types a sentence, which is passed for pre-processing to the Safar framework, which processes it and returns the corresponding lemma. Finally, the checker compares this lemma with all ACL lemmas. If the lemma of the input sentence matches any of the existing lemmas, the vocabulary of the sentence is considered to be covered by the ACL vocabulary, otherwise the words are rejected by the ACL vocabulary checker.

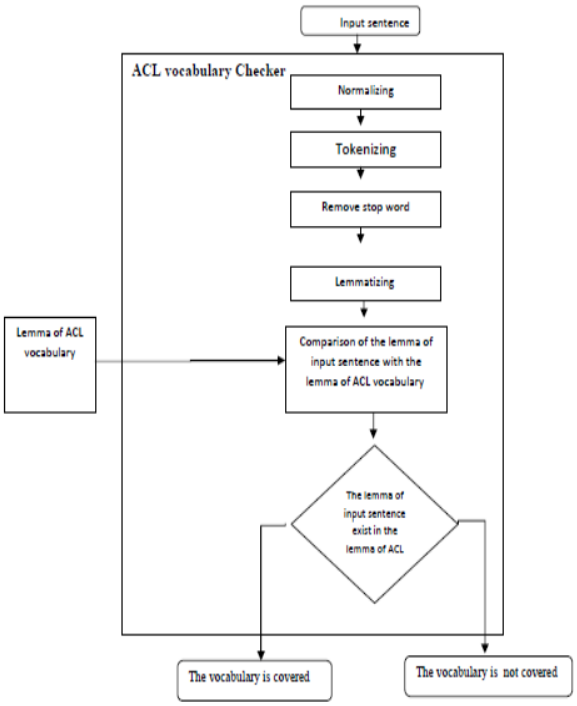

Figure 6: Overview of the ACL vocabulary checker architecture

We show next the output of three executions of the ACL vocabulary checker. In the first one, the system is given the sentence"هذه ممسحة كبيرة." /"This is a big mop.". As illustrated in Figure 7, the system performs the ACL vocabulary checking. For this first example, the sentence is covered by the ACL vocabulary and it outputs the input sentence, input sentence without stop words and lemma.

```
run:
The sentence: هذه ممسحة كبيرة.
sentances without stopwords:ممسحة كبيرة
The lemma of the words:
مِمْسَحٌ
كَبِيرٌ
 All words are covered by the ACL vocabulary
BUILD SUCCESSFUL (total time: ٣٣ seconds)
```

Figure 7: Analysis vocabulary of a sentence all vocabulary covered by ACL vocabulary

In the second example, we consider the sentence" هذه موزه كبيرة." /"This is a big banana.". The same steps are performed, but for this example, some words don't match the ACL vocabulary as illustrated in Figure 8.

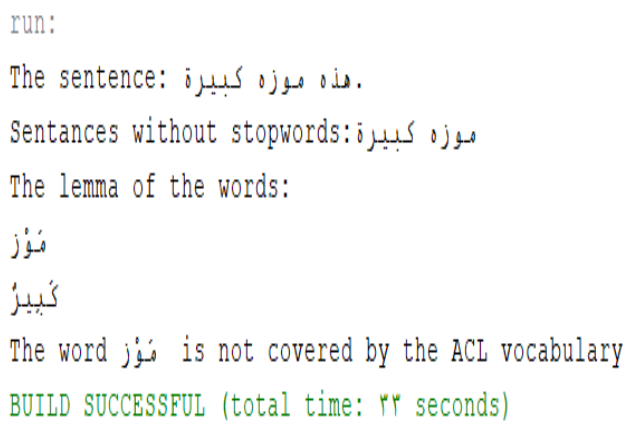

Figure 8: Analysis vocabulary of a sentence some vocabulary not covered by ACL vocabulary

In the third example, we consider the sentence"."السحاب يمطر" "Clouds raining.". The same steps are performed and as illustrated in Figure 9, all the words are rejected.

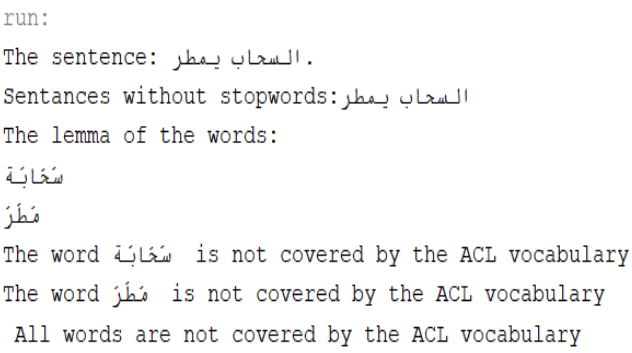

Figure9 : Analysis vocabulary of a sentence all vocabulary not covered by ACL vocabulary

V. CONCLUSION

In this paper, we presented the construction of linguistic resources for the development of the ACL semantic analyzer. In our approach, we first built the ACL lexicon from the ACL corpus and obtained all the vocabulary words found in the ACL corpus. We performed a pre-processing step on the sentences including tokenization, stop word removal, lemmatizing, and iteration removal. Then, we obtained an annotated ACL corpus with 354 sentences. To provide the semantic information for ACL sentences, we created an ACL ontology by adding 57.62% of ACL vocabularies to the Arabic ontology developed by Abouenour et al. [17]. In addition, we developed a tool called ACL vocabulary checker to verify the vocabulary of sentences against the ACL vocabulary. In the future, we plan to extract syntactic objects from the ACL grammar rules and generate semantic components to extract the meaning of sentences and formulates it using Conceptual Graphs formalism. This will open the space to develop a Machine translation tool based on ACL.

REFERENCE

[1] Palmer, M., Babko-Malaya, O., Bies, A., Diab, M., Maamouri, M., Mansouri, A., & Zaghouani, W. (2008, May). A pilot arabic propbank. In Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08).

[2] Shaalan, K. (2010). Rule-based approach in Arabic natural language processing. The International Journal on Information and Communication Technologies (IJICT), 3(3), 11-19.

[3] Maamouri, M., Bies, A., Buckwalter, T., & Mekki, W. (2004, September). The penn arabic treebank: Building a large-scale annotated arabic corpus. In NEMLAR conference on Arabic language resources and tools (Vol. 27, pp. 466-467).

[4] Habash, N., Dorr, B., &Monz, C. (2006). Challenges in building an Arabic-English GHMT system with SMT components. In Proceedings of the 11th Annual Conference of the European Association for Machine Translation (EAMT-2006), pp. 56 - 65.

[5] Hoyam Salah Elfahal Elebaed, Mohammed Kasbi, Mohammed Nasri,Karim Bouzoubaa, Khalid Tnaji. "Resources For Developing an Arabic Controlled Language". International Journal of Computer Science Trends and Technology (IJCST) V9(6): Page(49-61) Nov - Dec 2021. ISSN: 2347-8578.

[6] Schwitter, R. (2010, August). Controlled natural languages for knowledge representation. In Coling 2010: Posters (pp. 1113-1121).

[7] El Fahal, H. S., Nasri, M., Bouzoubaa, K.,&Kabbaj, A. (2019). Roadmap for an ArabicControlled Language. International Journal ofInformation Technology and Language Studies, 3(3).

[8] Black, W., Elkateb, S., Rodriguez, H., Alkhalifa, M., Vossen, P., Pease, A., & Fellbaum, C. (2006, January). Introducing the Arabic wordnet project. In Proceedings of the third international WordNet conference (pp. 295-300). Korea: Jeju.

[9] Mousser, J. (2010, May). A Large Coverage Verb Taxonomy for Arabic. In LREC.

[10] Mousser, J. (2011). Classifying Arabic verbs using sibling classes. In Proceedings of the Ninth International Conference on Computational Semantics (IWCS 2011).

[11] Kabbaj Adil. 2009. An overview of Amine. In: Pascal Hitzler, Henrik Scharfe, editors, Conceptual Structures in practice, pages 321-347.

[12] Nasri, M., ABOUENOUR, L., KABBAJ, A., & BOUZOUBAA, K. (2016). A novel approach for semantic analysis of Arabic texts using an Arabic ontology and Conceptual Graphs.

[13] Buitelaar, P. (2005). Human Language Technology for the Semantic Web.

[14] Beseiso, M., Ahmad, A. R., & Ismail, R. (2010). A Survey of Arabic language Support in Semantic web. International Journal of Computer Applications, 9(1), 35-40.

[15] Matuszek, C., Witbrock, M., Cabral, J., & DeOliveira, J. (2006). An introduction to the syntax and content of Cyc. UMBC Computer Science and Electrical Engineering Department Collection.

[16] Etzioni, O., Cafarella, M., Downey, D., Kok, S., Popescu, A. M., Shaked, T., ... & Yates, A. (2004, May). Web-scale information extraction in knowitall: (preliminary results). In Proceedings of the 13th international conference on World Wide Web (pp. 100-110).

[17] Abouenour, L., Nasri, M., Bouzoubaa, K., Kabbaj, A., & Rosso, P. (2014). Construction of an ontology for intelligent Arabic QA systems leveraging the Conceptual Graphs representation. Journal of Intelligent & Fuzzy Systems, 27(6), 2869-2881.

[18] Bouzoubaa, K., Jaafar, Y., Namly, D., Tachicart, R., Tajmout, R., Khamar, H., ... & Yousfi, A. (2021, April). A description and demonstration of SAFAR framework. In Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations (pp. 127-134).

[19] Pasha, A., Al-Badrashiny, M., Diab, M. T., El Kholy, A., Eskander, R., Habash, N., ... & Roth, R. (2014, May). Madamira: A fast, comprehensive tool for morphological analysis and disambiguation of arabic. In Lrec (Vol. 14, No. 2014, pp. 1094-1101).

[20] Souteh, Younes and Karim Bouzoubaa (2011). “SAFAR platform and its morphological layer”.3653 In: Proceeding of the Eleventh Conference on Language Engineering ESOLEC, pp. 14–153654 (cit. on pp. 29, 91, 96).

[21] Hoard, J. E., Wojcik, R. H., and Holzhauser, K. C. An Automated Grammar and Style Checker for Writers of Simplified English. In O'Brian, P. and Williams, N. (eds.), Computers and Writing: State of the Art. pp. 278-296, Intellect Books, Oxford, 1992.

[22] Nyberg, E. H., & Mitamura, T. (1996, March). Controlled language and knowledge-based machine translation: Principles and practice. In Proceedings of the first international workshop on controlled language applications (pp. 74-83).

[23] Hayes, P., Maxwell, S., &Schmandt, L. (1996, March). Controlled English advantages for translated and original English documents. In Proceedings of CLAW (pp. 84-92).