International Journal of Information Technology and Language Studies (IJITLS)

Roadmap for an Arabic Controlled Language

Hoyam Salah El Fahal[1], **Mohammed Nasri**[2], **Karim Bouzoubaa**[3], and **Adil Kabbaj**[4]
hoyam090@hotmail.com; mohammed.nasri@gmail.com; karim.bouzoubaa@emi.ac.ma; adil_kabbaj@yahoo.fr

[1] Faculty of Computer Science and Information Technology, Sudan University for Science and Technology, Khartoum, Sudan
[2] Laboratory of Science and Technology for the Engineer, National School for Applied Sciences, Khouribga, Sultan Moulay Slimane University, Morocco
[3] Computer Science Department, Mohammadia School of Engineers, Mohammed V University, Rabat, Morocco
[4] Computer Science Department, National Institute of Statistics and Applied Economics, Rabat, Morocco

Abstract. Controlled Natural Languages or CNLs are artificial subsets of natural languages that aim to make communication clearer and more precise. In general, CNLs are used in communication between humans or with computers and, particularly, when clarity and unambiguity are required. Existing CNLs have been developed to be exploited in many applications such as technical documentation, machine translation or database query language. So far, many CNLs have been developed for Western languages, especially English, but no concrete CNL has yet been proposed for Arabic even with the increasing number of Arabic Internet users in the last two decades. In this paper, we propose a roadmap for developing an Arabic CNL to provide new kind and advanced natural language services for Arabic people. Methodologically speaking, we review the most important existing CNLs in English and other languages helping us knowing some statistics related to the vocabulary size and the number of grammar rules that could help in designing the new CNL. This paper proposes two major approaches; one relies on leveraging on already-built CNLs, whereas the other consists in starting from scratch. The survey of Arabic NLP challenges along the available resources and tools allowed us to favor the second approach as the basis for the proposed roadmap.

Keywords: *Natural Language Processing; Controlled Natural Language; Arabic CNL; Arabic tools and resources; Arabic semantics.*

1. Introduction

Human-Computer Collaboration is the field that studies how Humans and computer agents work together to achieve a shared common goal. Computers can aid in accessing and recording information, and even analyzing collected information and making decisions (Zaroukian, 2016). To reach this shared goal, humans and computers need to communicate and exchange knowledge. Different ways can be used for this purpose such as the command line interface and the graphical user interface. Among them, the easiest way for humans is natural language. However, natural language is so large, complex and ambiguous that computers cannot easily and fully understand humans. Consequently, there is still a gap between computers and humans to understand each's one knowledge. To reduce this problem, some researchers have suggested, among other solutions, using formal languages called Controlled Natural Languages (CNLs).

CNLs are subsets of natural languages obtained by restricting their grammar and vocabulary (Fuchs et al., 2008). Humans need Controlled Languages for facilitating language learning, eliminating the need for frequent translation, streamlining translation and enhancing comprehensibility (Muegge, 2009). Over the last decade or so, a number of CNLs have been designed and used for specification purposes (Schwitter, 2002), knowledge acquisition (Gao, 2018), knowledge representation (Schwitter, 2010), and as interface languages to the Semantic Web (Kuhn et al., 2008).

The main advantage of CNLs over natural languages is that CNLs are both a natural and formal means of communication, making it adaptable for human-human as well as human-machine communication. Consequently, CNLs have been adopted in several fields, for instance in technical documentation (Fuchs, 1996), machine translation (Nyberg, 1996 and Mitamura, 1999), as a specification language (Schwitter, 2002) or as database query language (Androutsopoulos et al., 1995).

CNLs are not only used to facilitate the exchange of knowledge between humans and computers, they can also be used to improve communication among non-native speakers or to improve translation (Calderón, 2015).

Furthermore, some CNLs have been combined with other systems/tools to bolster some of their weaknesses or to let them provide new features; for instance, the Amine AI platform (Kabbaj, 2006) integrated the ACE CNL (Nasri et al., 2011). Indeed, this integration endowed the platform with an alternative and easy means (at least for humans) to represent knowledge. Such a combination of CNL and AI platform helps the platform i) provide a natural-language-based means of communication, and ii) be involved in other natural-language-based fields such as Machine Translation, QA or IR.

CNLs have already been developed for few languages such as English (Fuchs et al., 2008 and Schwitter et al., 2004), Spanish (Calderón, 2015) and German (Höfler, 2010). However, and to the best of our knowledge, no CNL has yet been developed for the Arabic language, even with the increasing growth of the Arabic digital content (the number of Arabic speaking internet user has increased by 8917.3% in the last nineteen years[1]) and the growing need of more services and interaction between Arabic speaking people and computers. Therefore, developing an Arabic controlled language will help in the design of new kind and advanced natural language services for Arabic people such as the improvement of current Arabic machine translation systems, more natural interfaces to the semantic web or more accurate named entity recognition systems. For instance, Shaalan (2014) acknowledges that the design of a new Arabic controlled language will increase the more accurate identification of named entities.

In this paper, we highlight the importance of CNLs and outline some existing CNLs and their uses in real applications and systems. The remainder of this paper is organized as follows: Section 2 presents the related works of existing CNLs and their uses. Section 3 ends up with our proposition for a new Arabic CNL and Section 4 concludes this paper.

2. Literature Review

CNLs are engineered languages that use a selection of the vocabulary, morphological forms, grammatical constructions, semantic interpretations, and pragmatics found in a natural language (Wyner, 2009). Various such languages exist today, at different stages of maturity. Pool (Pool, 2006)counts 41 projects that define controlled subsets of English, Esperanto, French, German, Greek, Japanese, Mandarin, Spanish and Swedish. Unsurprisingly most CNLs are based on English (Kuhn, 2014). They facilitate human-human communication (e.g. translation (Calderón, 2015) (Mitamura, 1999), technical documentation (Van, 1998)or even for medical literature (Kahn et al., 2014)) and human-machine communication (e.g. interfaces with databases (Wyner, 2009)or automated inference engines). For example, in the case of human-machine communication, CNLs enable casual users to use formal query languages such as SQL or SPARQL (Crego et al., 2016). Historically speaking, CNLs have also been previously used in the context of expert systems in order to make knowledge acquisition and maintenance more user-friendly (Kuhn, 2009) or in the context of knowledge representation (Gao, 2018 and Schwitter, 2010).

Our literature review is structured in two parts; the first one focuses on the most important English CNLs, whereas the second one is dedicated to other languages. We consider that the most important English CNLs are those that early appeared and are still in use. These CNLs will be presented in a chronological order (from the oldest to the newest). We end this section with a summarization of the most important features that will help us in designing a new CNL for another language such as Arabic.

[1] Extracted in June 5th, 2019from: https://www.internetworldstats.com/stats7.htm

2.1CNLs for the English language

As per Kuhn survey (Kuhn, 2014), the top 10 CNLs that were created early but are still used are: Basic English, Special English, Plain Language, CAA Phraseology, SMART Plain English, Standard Language, RuleSpeak, Attempto Controlled English, Gellish English, and EasyEnglish. The rest of this chapter gives a brief detail of each one.

Basic English (Ogden, 1930) is the first controlled version of the English language introduced by Ogden in 1930 and designed to improve the communication among people around the globe. It aimed to achieve this by reducing the lexicon, the syntax and the grammar of Standard English into ten writing rules and a vocabulary core made-up of 850 root words. Here is an exemplary excerpt of a text in Basic English:

- "It was his view that in another hundred years Britain will be a second-rate power."

Special English (voice of America, 2009) is a simplified English introduced in 1959 and used for radio and television news[2]. Special English has been used later in the SEASPEAK project (Stevens et al., 1983) which is another CNL designed to facilitate communication between ships, whose captains' native languages differ. This project has now been formalized as SMCP – Standard Marine Communication Phrases (Trenkner, 2005). We cite in what follows to examples of Special English[3]:

- "Winter weather in Washington, D.C. can be really windy. And wind messes up my hair. It is really windy today. Look at my hair. Will it be windy this weekend? I'll listen to the news. I am tired of my untidy hair."

- "Are you tired of your untidy hair?"

Plain Language (SEC, 1998), also known as Plain English was introduced by the US government and other organizations about 1970 as an initiative to make official documents easier to understand and less bureaucratic. The purpose was also to provide an easy language for investors, brokers, investment advisers, lawyers and so on. The Plain Language team highlights many problems that make a standard text ambiguous and less clear such as using long sentences, passive voice, weak verbs or unnecessary details and offer, against, ways to fix them. Two examples of Plain Language are introduced below:

- "You should rely only on the information contained in this document or that we have referred you to. We have not authorized anyone to provide you with information that is different.".

- "The Board might approve these investments in advance."

CAA Phraseology (Authority, 2011), ICAO Phraseology (Phraseology), and FAA Air Traffic Control Phraseology (FAA, 2015) are languages for air traffic control respectively introduced by the CCA (Civil Aviation Authority), ICAO (International Civil Aviation Organisation) and the FAA (Federal Aviation Administration) in the 1980s and are very similar to each other. These languages gave birth to the Radiotelephony Communication System for Pilots, called AIRSPEAK (Robertson, 1987 and 2008). This is an example of Phraseology CNL:

- "Take the second turning on the left."

SMART Plain English[4], or SMART's Plain English Program (PEP) is a CNL developed and used at SMART Communications, Inc., since 1980 (Kuhn, 2014). SMART Plain English influenced many others CNLs such as Bull Global English (SC, 1994), Controlled English at Clark (Adriaens et al., 1992), Controlled English at Rockwell (Adriaens et al., 1992), Nortel Standard English (Smart, 2006) and SMART Controlled English (Smart, 2006).

[2] https://learningenglish.voanews.com/

[3] Extractedfrom a conversationat: https://learningenglish.voanews.com/a/lets-learn-english-lesson-39-its-unbelieveable/3598920.html

[4] http://www.smartny.com/plainenglish.htm

SMART provides some tools[5] that can be used for several CNL, among then PEP. We cite for example MAXit Checker[6], which can be used to create compliant documents. Here are two examples of SMART Plain English sentences:

- "How to enroll in the donate life registry"

- "You can enroll at either the offices of the Department of Motor Vehicles (DMV) or the Department of Health (DOH)"

Standard Language (Rychtyckyj, 2005), also known as SLANG, has been introduced by Ford Company as a standard format for writing process descriptions. The goal of the SLANG was to develop a clear and consistent means of communicating for a so-called process build instructions between various engineering functions. The use of SLANG across Ford Company has eliminated almost all ambiguity in instructions and has created a standard format for writing documents.

Within the company, a single vehicle may require thousands of so called process sheets; each sheet contains the detailed instructions needed to build a portion of a vehicle as well as its associated part and tooling information. SLANG is extremely useful as it allows an engineer to write clear and concise assembly instructions that are unambiguous and machine-readable.

SLANG has been used in conjunction with a Machine Translation (Rychtyckyj, 2002 and 2005) to allow translation of the manufacturing processes for plants in other countries. Here are two examples of SLANG:

- "APPLY GREASE TO RUBBER O-RING AND CORE OPENING"

- "INSERT HEATER ASSEMBLY INTO RIGHT REAR CORE PLUGHOSE"

RuleSpeak (Ross, 1996, 2009a and 2009b) is a CNL for business rules. It is defined in (Ross, 2009a) as a set of practical guidelines for 1) expressing Business Rules in clear, unambiguous, well-structured English, 2) improving communication about Business Rules among business people, business analysts, and IT professionals, 3) bridging the gap between the language of business policies and legal obligations, and IT specifications oriented to system design and implementation, 4) avoiding common pitfalls associated with expressing guidance, and for 5) retaining product/service know-how in a manageable, reusable form. These two sentences are examples of RuleSpeak:

- "An item may be returned if some proof of purchase is provided."

- "A person of any age may hold a bank account."

Attempto Controlled English (Fuchs et al. 1996) or simply ACE is a subset of English introduced by Fuchs and Schwitter in 1996 as a language for software specifications. Its focus shifted afterward towards knowledge representation and the Semantic Web. Attempto team has been developing many tools around ACE including APE (Attempto Parsing Engine), a tool that translates ACE text into first-order logic via DRS (Discourse Representation Structures) (Kamp et al., 2013). ACE is defined by a small set of construction rules (Fuchs et al., 2008) that describe its syntax. Its vocabulary consists of predefined function words, some predefined fixed phrases (there is, it is false that, etc.), and content words.ACE allows representing specifications in simple or composite sentences (coordination, subordination, quantification, and negation). The most notable features of ACE include complex noun phrases, plurals, anaphoric references, subordinated clauses, modality, and questions. These are two exemplary ACE sentences:

- "A customer owns a card that is invalid or that is damaged."

- "Every continent that is not Antarctica contains at least 2 countries."

[5] The othertoolscanbefoundat: http://www.smartny.com/

[6] http://www.smartny.com/maxit.htm

Gellish (Renssen, 2005 and 2013) is a knowledge representation language and ontology appeared in 1998. The ontology defines a rich and extensible semantics in natural language terminology, expressed in Gellish English itself. This ontology is equivalent to a data model of over 20.000 entity types, attribute types and relationship types. Gellish is fact oriented, which means, a Gellish object is a fact. Each atomic fact is expressed as a relation between two objects. These are two sentences of Gellish objects:

- “John is performer of action#1”
- “action#1 is classified as maintenance”

Easy English (Bernth, 1997) is a controlled English introduced in 1997 as part of IBM's internal SGML editing environment that helps writers produce clearer and simpler English. IBM provided a tool that helps pointing out ambiguity and where appropriate, makes suggestions for rephrasing. That tool allows in addition performing some standard grammar checking. Here are two exemplary Easy English sentences:

- “Different system users may operate on different objects by using the same application program.”
- “It is the number defined in the result field definition or the file.”

2.2 CNLs for other languages

CLG (Höfler, 2010) is a German CNL for the representation of legal norms contained in Swiss statutes and regulations to the development of knowledge-based legal information systems. CLG restricts the syntax and semantics of Swiss legal language to prevent instances of ambiguity arising from constructions that either have more than one syntactic analysis (syntactic ambiguity) or whose syntactic analysis can be mapped onto more than one non-equivalent logical structure (semantic ambiguity). The first version CLG 1.0 (published in 2010) provides the basic syntactic and semantic inventory to express simple norms (obligations, permissions, prohibitions; including norms stating duties and responsibilities). It comprises roughly two dozens construction and interpretation rules that deal with phenomena such as attachment ambiguities, plural ambiguities, scope ambiguities, lexical ambiguities, referential ambiguities and functional ambiguities due to the relatively free German word order.

INAUT (Haralambous, 2014) is a French CNL dedicated to collaborative update of a knowledge base on maritime navigation and to automatic generation of coast pilot books of the French National Hydrographic and Oceanographic Service SHOM. The INAUT is a controlled language with a rather large vocabulary but with a simple syntax.

In 2015, Miyata et al. (Miyata et al., 2015) reported on experiments to test the effectiveness of controlled language rules on texts from Japanese municipal websites to Improve Machine Translatability of Municipal Documents. They compiled a set of rules by trial and error, systematically rewriting Japanese source texts and analyzing the machine translation outputs. They employed native English speakers with little knowledge of Japanese as human evaluators to test the understandability and accuracy of the English machine translated text.

CAL (Elazhary, 2016) is a first Controlled Arabic Language for authoring ontologies using Arabic words only. CAL is accompanied by a tool that allows translating CAL statements to OWL. Detailed error messages are also generated in Arabic. Besides, CAL is based on and has a similar syntax as the Rabbit CNL (Hart et al., 2008), and thus, statements can be easily translated between the two languages, allowing the cooperation of Arabic-speaking and English-speaking domain experts, and ontology engineers in ontology authoring and validation, a capability which is absent in other CNLs.

2.3 Summary

We notice from this review that CNLs are more numerous and more developed for English than for other languages. We also notice that there is almost no concrete one for Arabic.

Among the mentioned English CNLs, ACE can be considered as one of the best since it is: 1) a well-known[7] and general-purpose English CNL, 2) designed for human-human and human-machine communication, 3) well developed[8], and 4) mature and well cited[9].ACE has been considered in other works (Nasri, 2016) too as one of the best English CNL.

As a starting point toward designing an Arabic CNL, ACE can be enough as reference. The following section introduces some ideas for a new CNLs as well as our roadmap for creating one.

3. Ideas for a new Arabic CNL

3.1 Motivations and potential perspectives

With the rapid growth of Arabic digital content, there is an increasing need for more services and tools dedicated to Arabic citizens. Considering Arabic at large, many researchers are making efforts to provide such tools such as Arabic morphological analyzers (Smrž, 2007), lemmatizes (Mubarak, 2017), sentiment analysis systems (Alhumoud, 2015) and so on.

One other possibility is to develop a new CNL for Arabic helping the community develop more services and applications for Arabic people such as the improvement of current Arabic machine translation systems, more natural interfaces to the semantic web or more accurate named entity recognition systems.

This is basically the case for syntactic as well as semantic analysis tools. Indeed, some CNLs have been used for semantic analysis such as the APE tool developed by the Attempto group to convert any English text respecting the ACE CNL into the DRS semantic structure (Kamp et al., 2013). ACE CNL has also been used to be converted into the Conceptual Graph semantic formalism (Sowa, 1983). The same approach could be adopted to allow a semantic analysis of Arabic through a CNL. In addition, if such a semantic analyzer is provided, it will certainly be the basis for many other semantic-based works such as QA, IR or even Machine Translation.

As previously noted and to the best of our knowledge no Arabic CNL exists. To develop such CNL dedicated to Arabic, one might take advantage of the experiences of existing ones in terms of technical and scoping constraints. Therefore, the design of our new ACL (Arabic Controlled Language) should respect the following:

- It should cover a large wide of Arabic vocabulary so that users feel free to express knowledge easily. The size of such vocabulary should be neighboring to that of one of the best existing English CNLs.

- As a starting point, the ACL should cover grammar rules enough for an intermediate level of Arabic.

- The ACL should respect the Arabic grammar to cover many kinds of its POS such as function words, nouns, verbs, particles, adjectives, and adverbs and to cover also both verbal and nominal sentences

- Such ACL should allow not only simple sentences but also composite ones such as conjunction, modality, interrogative, etc.

- CNLs are generally developed for one of three purposes: business rule redactions, as semantic web language or for general purposes. Our target ACL is not limited to a specific scope or domain but rather will be dedicated to general purposes.

[7] The first ranked CNL when searching for « Controlled Natural English » in Google Scholar, operation made on June 5th, 2019

[8] Lexicon, API and many tools are available at: http://attempto.ifi.uzh.ch/site/resources/

[9] According to Google Scholar, it has been cited 212 times, so far, June 5th, 2019.

- In addition to vocabulary and grammar, and in order to provide an exploitable and useful ACL, Arabic language processing tools should be available such as sentence checking, syntactic analysis or even semantic analysis.

3.2 Possible approaches

Our literature review revealed that existing CNLs were developed using different methods. Therefore, there is not one single approach for such development. From the specifications we outlined above, ACL might be developed using two major approaches; one relies on one or more already-built CNLs whereas the other consists of starting from scratch.

Concerning the first method, and as mentioned above, ACE is one of the best English CNL and can be used as a reference. It is, therefore, a good choice to start with where the approach will consist in considering ACE and translating all or a part of it. This means translating respectively, when possible, grammar rules and vocabulary to Arabic grammar rules and Arabic vocabulary.

The second approach consists of constructing a new ACL from scratch without benefitting from others' works. This means identifying a subset of Arabic grammar rules and lexicon that is representative of the domain in hand. Of course, the subset should be homogeneous avoiding both ambiguity and complexity.

Each approach has its advantages and disadvantages; the first one is interesting in that it allows benefitting from others' experiences and avoids starting from scratch or making the same mistakes of the others. However, implementing it is not that easy because i) Arabic grammar is different from English and ii) the tools used for other CNLs are not necessarily reusable for the Arabic language.

The second approach is cleaner since it is 100 percent specific to Arabic, and everything is thought for this language, but this also means doing all the work ourselves; identifying the subset of the language, formulating the rules, building the lexicon, choosing tools and performing implementations.

3.3 Difficulties and challenges

Works on Arabic NLP are a bit complicated in comparison to other languages such as English or Western languages. This is due on one hand to the language itself and on the other hand to the limited amount of tools and resources. We illustrate below two fundamental linguistic differences between Arabic and English and then expose the most important tools and resources with a focus on those needed in the context of controlled languages.

First, Arabic offers the ability to paste elements such as articles, prepositions, and pronouns to the adjectives, nouns, verbs, and particles they relate to. Therefore, an Arabic word can sometimes correspond to a whole sentence in another language. For example, the Arabic word "فسيكفيكهم" corresponds in English to the sentence "Then he (Allah) will suffice you against them.". In addition, it is sometimes difficult to distinguish between a proclitic or enclitic and a character of the word in hand. For example, the character "و" (w) in the word "وصل" (arrived) is part of that word whereas in the word "ومر" (and is passed), it is a proclitic.

On another hand, Arabic is characterized by the absence of short vowels called "diacritics –الحركات " in most written texts. Indeed, Arabic short vowels are not letters of the alphabet, they are diacritic signs that are added to consonants and play the same role as vowels in other languages. The writings in Arabic are generally not diacritical and it is up to the reader to guess the diacritics of the texts at the reading time. On the other hand, religious texts and some school textbooks are entirely diacritical. Other resources such as journalistic texts may be partially diacritical. The diacritics added in these writings are used to remove morphological, syntactic and sometimes semantic ambiguities. Indeed, i) diacritics are very important to remove syntactic ambiguity, they are associated with the last letter of a nominal word and they mark the case, ii) they help identify the syntactic functions of words in a sentence, iii) diacritics can also be assigned to other letters to remove morphological and semantic ambiguities, this characteristic can be illustrated in the word: "كتب" which, without diacritics, can mean "books –كتب " in the sentence " كتب التلميذ في الخزانة" (the student's books are in the cupboard) or "he wrote – كتب" in the sentence " كتب التلميذ الدرس" (the student wrote the lesson).

Concerning Arabic resources, and to the best of our knowledge, there is no resource or corpus specifically dedicated to Controlled Languages. A survey of other available Arabic corpora can be found in (Zaghouani, 2017). For the tools side, controlled languages may be in need of morphological analyzers, syntactic parsers, semantic analyzers or even a combination of these tools. We list below the most important existing works targeting the Arabic language.

The role of a Morphological analyzer is to descript the minimal constituents of a word in a text. These analyzers main objective is to generate all possible analyses of the words out of their contexts such as stem, root, pattern, affixes, etc. Several researchers have investigated different approaches to Arabic morphological analysis such as: Buckwalter's Morphological Analyzer BAMA (Buckwalter2002), ElixirMF (Smrž, 2007), SALMA (Sawalha et al., 2013), Al-Khalil (Boudchiche et al., 2017 and Boudlal et al., 2010), Sarf (Elghazaly, 2015), SAMA (Zaraket, 2012), MORPH2 (Kammoun et al., 2010), MADAMIRA (Pasha et al., 2014), etc. Authors of (Jaafar et al. 2016) evaluated the performance of many morphological tools using a reference corpus. Boudchiche et al. (Boudchiche et al., 2017) also evaluated the performance of Alkhalil2, Alkhalil Morpho Sys, BAMA, and SAMA using a large corpus of more than 72 million diacritized words. Alkhalil2 analyzer was able to analyze 99.31% of the words against only 90.18% for SAMA analyzer and a lower rate for the other two analyzers.

The aim of a syntactic parser is analyzing a string of symbols conforming to the rules of grammar. Key syntactic parsers for Arabic text are: Stanford (Klein et al., 2003), ATKS (Microsoft 2017) and Farasa (Abdelali et al., 2016). Stanford Parser generates a Treebank parse tree for the input sentence. Farasa Dependency parser is based on SVM-rank using linear kernels that use a variety of features and lexicons to rank possible segmentations of a word. However, the existing Arabic parsers are not conforming to all rules of the Arabic grammar. Authors of (Jaafar and Bouzoubaa 2018) performed a benchmark between the Standford and the ATKS Arabic parsers using the OntoNotes corpus (Pradhan et al. 2007).

Concerning Semantic analyzers, there is only two works devoted to semantically analyze Arabic sentences. The first approach presented by SamehAlansary (Alansary et al., 2013), which introduced the UNL framework, is able to analyze automatically natural languages into their abstract semantic meanings, with the aim of finding the common denominator between all languages. Moreover, the data is exportable in several different formats. The second approach, introduced by Nasri et al. (Nasri et al., 2013), is made-up of two modules. The first one consists in building an Arabic ontology that provides, in addition to its entries (nouns, verbs, adjectives, prepositions, etc.), semantic information related to the various meanings of each verb depending on the situation this verb is used in. The second module is based on the ontology (provided by module 1) as well as some linguistic tools (Stanford Parser and Al Khalil morphological analyzer) and, for each parsed sentence; this module extracts the main verb and the syntactic structure via a syntactic and morphological analysis. The analysis result is combined with the semantic information provided by the already-built ontology and used by this module to deduct the meaning of the sentence. In their approach, Nasri et al. formulate the semantic information (extracted from the text) by means of the Conceptual Graph formalism.

3.4 Discussion

Recall that we have proposed two major approaches for creating a new ACL, the first relies on one or more already-built CNLs whereas the second is to start from scratch. At the first stage, all efforts and works will be oriented toward language analysis and comprehension, the language generation will be the subject of other future works.

According to the Arabic NLP challenges and the available tools and resources, we are currently involved in developing a new general-purpose ACL, which is based on the second approach. This means we are currently identifying a subset of Arabic, with the size of lexicon neighboring the best existing English CNLs. As previously mentioned, we may consider ACE as a reference for this purpose. Table 1 highlights some details about the vocabulary of ACE.

Category	Vocabulary size
Nouns	21,700
Verbs	12,700
Adjectives	17,500
Adverbs	2,300
Pronoun	80
Preposition	50
Summary	54,300

Table 1. Vocabulary size of ACE per category[10]

From Table 1, and as a starting point, we can consider that a vocabulary of 50,000 entries would suffice for our ACL. To do this, and since there is no available Arabic CNL corpus, we will begin by extracting the rules and vocabulary used in Arabic textbooks of primary school. Indeed, this type of textbooks contains both verbal and nominal sentences, simple and composite sentences (conjunction, modality, interrogative and so on) and offers many kinds of POS (function words, nouns, verbs, particles, adjectives, and adverbs).

At the same time, we are developing benchmarks allowing us to select among the mentioned available ones, the best Arabic morphological analyzer and the best Arabic syntactic parser. These two tools will be used in order to analyze the developed ACL corpus to extract the syntactic/grammar rules that will constitute our ACL.

Later, we aim at developing the corresponding semantic analyzer aiming to extract the meaning conveyed by a text and formulate it in knowledge representation formalism. This can be useful not only for textbook text analysis itself, but also in many concrete semantic-based applications such as IR, QA, or even Machine Translation.

4. Conclusion

At first glance, we showed how important is the CNLs for NLP. We showed also that there is, so far, no existing CNL for many languages such as Arabic. To propose a roadmap for the development of a new Arabic CNL, it was necessary to make a survey of the available CNLs and have an idea about languages of the existing CNLs, statistics about the size of rules and vocabulary as well as the applications where it could be applied. Then, the paper proposed two major approaches for such development; one relies on leveraging on already-built CNLs whereas the other consists of starting from scratch. The survey of Arabic NLP challenges along the available resources and tools allowed us to favor the second approach as the basis for the proposed roadmap.

Therefore, our findings about the existing Arabic CNL, the existing Arabic resources and tools imply to choose as next steps of our project (i) compile an Arabic corpus from textbooks; (ii) benchmark the many existing morphological analyzers to select the best one; and (iii) benchmark the existing syntactic parsers. These steps will allow to extract the grammar rules from the compiled corpus, extract the vocabulary using the corpus and the selected morphological analyzer and exploit the syntactic tool to parse all sentences from the corpus. Finally, the last step will be to develop a corresponding semantic analyzer.

References

Abdelali, Ahmed, et al. "Farasa: A fast and furious segmenter for Arabic." Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations. 2016.

Abouenour, Lahsen, et al. "Building an Arabic morphological analyzer as part of an open Arabic NLP platform." Workshop on HLT and NLP within the Arabic world: Arabic Language and local languages

[10] Version 6.7 of ACE lexicon, available at: https://raw.githubusercontent.com/Attempto/Clex/master/clex_lexicon.pl

processing Status Updates and Prospects At the 6th Language Resources and Evaluation Conference (LREC’08). 2008.

Adriaens, Geert, and Dirk Schreors. "From COGRAM to ALCOGRAM: Toward a controlled English grammar checker." *Proceedings of the 14th conference on Computational linguistics-Volume 2*. Association for Computational Linguistics, 1992.

Alansary, Sameh, MagdyNagi, and NohaAdly. "A suite of tools for Arabic natural language processing: A UNL approach." *2013 1st International Conference on Communications, Signal Processing, and their Applications (ICCSPA)*. IEEE, 2013.

Alhumoud, Sarah O., et al. "Survey on arabic sentiment analysis in twitter." *International Science Index* 9.1 (2015): 364-368.

Androutsopoulos, Ion, Graeme D. Ritchie, and Peter Thanisch. "Natural language interfaces to databases–an introduction." Natural language engineering 1.1 (1995): 29-81.

Authority, Civil Aviation. "CAP 413: Radiotelephony Manual." (2010).

Bernth, Arendse. "EasyEnglish: a tool for improving document quality." *Proceedings of the fifth conference on Applied natural language processing*. Association for Computational Linguistics, 1997.

Boudchiche, Mohamed, et al. "AlKhalilMorpho Sys 2: A robust Arabic morpho-syntactic analyzer." *Journal of King SaudUniversity-Computer and Information Sciences* 29.2 (2017): 141-146.

Boudlal, Abderrahim, et al. "Alkhalilmorpho sys1: A morphosyntactic analysis system for arabic texts." International Arabconference on information technology. Benghazi Libya, 2010.

Buckwalter, T. (2002a). Arabic morphology analysis. Retrieved April23, 2015, from QAMUS: http://www.qamus.org/morphology.htm.

Calderón, Sebastián León. "Building a Controlled Natural Language Framework for Real-time Machine Translation." Revista de LenguasModernas 23 (2015).

Crego, Josep, et al. "Systran's pure neural machine translation systems." *arXiv preprint arXiv:1610.05540* (2016).

Elazhary, Hanan. "CAL: A controlled Arabic language for authoring ontologies." *Arabian Journal for Science and Engineering* 41.8 (2016): 2911-2926.

Elghazaly, T., and A. M. Maabid. "Assessing and Evaluating Arabic Morphological Analyzers and Generators." Future Communication Technology and Engineering: Proceedings of the 2014 International Conference on Future Communication Technology and Engineering (FCTE 2014), Shenzhen, China, 16-17 November 2014. CRC Press, 2015.

FAA (Federal Aviation Administration) JO Order 7110.65W, Air Traffic Control

Fuchs, Norbert E., and Rolf Schwitter. "Attempto controlled english (ace)." *arXiv preprint cmp-lg/9603003* (1996).

Fuchs, Norbert E., KaarelKaljurand, and Tobias Kuhn. "Attempto controlled english for knowledge representation." *Reasoning Web*. Springer, Berlin, Heidelberg, 2008. 104-124.

Gao, Tiantian. "Achieving High Quality Knowledge Acquisition using Controlled Natural Language." *Technical Communications of the 33rd International Conference on Logic Programming (ICLP 2017)*. SchlossDagstuhl-Leibniz-ZentrumfuerInformatik, 2018.

Gridach, Mourad, and NoureddineChenfour. "Developing a new approach for arabic morphological analysis and generation." *arXiv preprint arXiv:1101.5494* (2011).

Haralambous, Yannis, Julie Sauvage-Vincent, and John Puentes. "INAUT, a controlled language for the French coast pilot books instructions nautiques." *International Workshop on Controlled Natural Language*. Springer, Cham, 2014.

Hart, Glen, Martina Johnson, and Catherine Dolbear. "Rabbit: Developing a control natural language for authoring ontologies." European Semantic Web Conference. Springer, Berlin, Heidelberg, 2008.

Höfler, Stefan, and Alexandra Bünzli. "Designing a controlled natural language for the representation of legal norms." *Second Workshop on Controlled Natural Languages*. 2010.

Y. Jaafar, K. Bouzoubaa, A. Yousfi, R. Tajmout, H. Khamar, "Improving Arabic Morphological Analyzers Benchmark", In The International Journal of Speech Technology (IJST), pp. 1-9, April 2016

Jaafar Y., Bouzoubaa K. (2018) "A New Tool for Benchmarking and Assessing Arabic Syntactic Parsers". In: Lachkar A., Bouzoubaa K., Mazroui A., Hamdani A., Lekhouaja A. (eds) Arabic Language Processing: From Theory to Practice. ICALP 2017. Communications in Computer and Information Science, vol 782. Springer, Cham

Kabbaj, Adil. "Development of intelligent systems and multi-agents systems with amine platform." International Conference on Conceptual Structures. Springer, Berlin, Heidelberg, 2006.

Kahn JM, Gould MK, Krishnan JA, Wilson KC, Au DH, Cooke CR, Douglas IS, Feemster LC, Mularski RA, Slatore CG, Wiener RS. "An official American thoracic society workshop report: developing performance measures from clinical practice guidelines.". ATS Ad Hoc Committee on the Development of Performance Measures from ATS Guidelines.Ann Am Thorac Soc. 2014 May;11(4):S186-95. doi: 10.1513/AnnalsATS.201403-106ST.

Kammoun, NouhaChaâben, Lamia HadrichBelguith, and Abdelmajid Ben Hamadou. "The MORPH2 new version: A robust morphological analyzer for Arabic texts." *JADT 2010: 10th International Conference on Statistical Analysis of Textual Data*. 2010.

Kamp, Hans, and Uwe Reyle. From discourse to logic: Introduction to modeltheoretic semantics of natural language, formal logic and discourse representation theory. Vol. 42. Springer Science & Business Media, 2013.

Klein, Dan, and Christopher D. Manning. "Accurate unlexicalized parsing." *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics-Volume 1*. Association for ComputationalLinguistics, 2003.

Kuhn, Tobias, and Rolf Schwitter. "Writing support for controlled natural languages." *Proceedings of the Australasianlanguagetechnology association workshop 2008*. 2008.

Kuhn, Tobias. "A survey and classification of controlled natural languages." *ComputationalLinguistics* 40.1 (2014): 121-170.

Kuhn, Tobias. *Controlled English for knowledge representation*. Diss. Doctoral thesis, Faculty of Economics, Business Administration and Information Technology of the University of Zurich, Switzerland, to appear, 2009.

Microsoft, "Arabic Toolkit Service (ATKS)," [Online]. Available: https://www.microsoft.com/en-us/research/project/arabic-toolkit-service-atks/. [Accessed 01 03 2017].

Mitamura, Teruko. "Controlled language for multilingual machine translation." *Proceedings of Machine Translation Summit VII, Singapore*. 1999.

Miyata, Rei, et al. "Japanese controlled language rules to improve machine translatability of municipal documents." *Proc. of MT Summit*. 2015.

Mubarak, Hamdy. "Build fast and accurate lemmatization for Arabic." *arXiv preprint arXiv:1710.06700* (2017).

Muegge, Uwe. "Controlled language-does my company need it?." *URL: www.tekom.de/artikel/artikel_2756 html* (2009).

Nasri, M., et al. "Toward a semantic analyzer for Arabic language." *22nd IBIMA* (2013).

Nasri, Mohammed, AdilKabbaj, and Karim Bouzoubaa. "Integration of the controlled language ace to the amine platform." *International Conference on Conceptual Structures*. Springer, Berlin, Heidelberg, 2011.

Nasri, Mohammed. Intégration d'une composante pour le traitement du langage naturel dans une plateforme pour les systèmes intelligents. Doctoral dissertation. Ecole Mohammadia d'Ingénieurs, 2016.

Nyberg, Eric H., and Teruko Mitamura. "Controlled language and knowledge-based machine translation: Principles and practice." *Proceedings of the first international workshop on controlled language applications*. Vol. 74. 1996.

Ogden, C. K. (1930). Basic English: A general introduction with rules and grammar.

Pasha, Arfath, et al. "Madamira: A fast, comprehensive tool for morphological analysis and disambiguation of arabic." *LREC*. Vol. 14. 2014.

Phraseology, ICAO ICAO Standard. "A Quick Reference Guide for Commercial Air Transport Pilots." ICAO Phraseology Ref. Guide: 1-19.

Pool, Jonathan. "Can controlled languages scale to the Web?." *International Workshop on Controlled Language Applications 5*. 2006.

S. S. Pradhan, E. Hovy, M. Marcus, M. Palmer, L. Ramshaw and R. Weischedel, "Ontonotes: A unified relational semantic representation," *International Journal of Semantic Computing,* vol. 1, no. 04, pp. 405-419, 2007.

Robertson, Fiona A. Airspeak. Pearson Education, 2008.

Robertson, Fiona A., and Edward Johnson. Airspeak. Radiotelephony communication for pilots. 1987.

Ross, R. G. "Rulespeak." *Business Rule Solutions, LLC* (1996).

Ross, Ronald G. "Basic RuleSpeak Guidelines." Do's and Don'ts in Expressing Natural-Language Business Rules in English (2009a).

Ross, Ronald G. "RuleSpeak Sentence Forms: Specifying Natural-Language Business Rules in English." *Business Rules Journal* 10.4 (2009b).

Rychtyckyj, Nestor. "An assessment of machine translation for vehicle assembly process planning at Ford motor company." *Conference of the Association for Machine Translation in the Americas*. Springer, Berlin, Heidelberg, 2002.

Rychtyckyj, Nestor. "Ergonomics analysis for vehicle assembly using artificial intelligence." *AI Magazine* 26.3 (2005): 41-41.

Sawalha, Majdi, Eric Atwell, and Mohammad AM Abushariah. "SALMA: standard Arabic language morphological analysis." *2013 1st International Conference on Communications, Signal Processing, and their Applications (ICCSPA)*. IEEE, 2013.

SC (Smart Communications Inc.). News from Smart Communications, Inc. In MT News International—Newsletter of the International Association for Machine Translation. 1994. Issue no. 7.

Schwitter, Rolf, and Marc Tilbrook. "Controlled natural language meets the semanticweb." *Proceedings of the AustralasianLanguageTechnology Workshop 2004*. 2004.

Schwitter, Rolf. "Controlled natural languages for knowledge representation." *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*. Association for ComputationalLinguistics, 2010.

Schwitter, Rolf. "English as a formal specification language." Proceedings. 13th International Workshop on Database and Expert Systems Applications. IEEE, 2002.

SEC (Securities and Exchange Commission). "A plain English handbook: How to create clear SEC disclosure documents." *US Securities and Exchange Commission, Washington, DC* (1998).

Shaalan, Khaled. "A survey of arabic named entity recognition and classification." Computational Linguistics 40, no. 2 (2014): 469-510.

Smart, John M. "SMART controlled English." *Proceedings of CLAW 2006* 9 (2006).

Smrž, Otakar. "Elixirfm: implementation of functional arabic morphology." *Proceedings of the 2007 workshop on computational approaches to Semitic languages: common issues and resources*. Association for ComputationalLinguistics, 2007.

Sowa, John F. "Conceptual structures: information processing in mind and machine." (1983).

Strevens, Peter, and Edward Johnson. "SEASPEAK: A project in applied linguistics, language engineering, and eventually ESP for sailors." *The ESP Journal* 2.2 (1983): 123-129.

Trenkner, Peter. "The IMO Standard Marine Communication Phrases–Refreshing memories to refresh motivation." *Proceedings of the IMLA 17th International Maritime English Conference*. 2005.

Van der Eijck, P. "Controlled languages in technical documentation." *Selected Papers from the Eight CLIN meeting*. 1998.

Van Renssen, A. "Gellish Formal English. Definition and Application of a Universal Information Modeling Language." (2013).

Van Renssen, Andries Simon Hendrik Paul. "Gellish: a generic extensible ontological language-design and application of a universal data structure." (2005).

Voice of America. 2009. VOA Special English Word Book: A List of Words Used in Special English Programs on Radio, Television, and the Internet,Washington, DC. Warren, David H. D. and Fernando

Wyner, Adam, et al. "On controlled natural languages: Properties and prospects." *International Workshop on Controlled Natural Language*. Springer, Berlin, Heidelberg, 2009.

Zaghouani, Wajdi. "Critical survey of the freely available Arabic corpora." *arXiv preprint arXiv:1702.07835* (2017).

Zaraket, Fadi, and JadMakhlouta. "Arabic morphological analyzer with agglutinative affix morphemes and fusional concatenation rules." *Proceedings of COLING 2012: Demonstration Papers* (2012): 517-526.

Zaroukian, Erin. "Human understanding of Controlled Natural Language in simulated tactical environments." 2016 IEEE International Multi-Disciplinary Conference on Cognitive Methods in Situation Awareness and Decision Support (CogSIMA). IEEE, 2016.