Journal of Computer Sciences and Applications, 2013, Vol. 1, No. 2, 17-22
Available online at http://pubs.sciepub.com/jcsa/1/2/1
© Science and Education Publishing
DOI:10.12691/jcsa-1-2-1

The Interactive Arabic Dictionary: Another Collaboratively Constructed Language Resource

Ghaida Rebdawi[*], Said Desouki, Nada Ghneim

Computer Science Department, HIAST (Higher Institute for Applied Sciences and Technology), Damascus, Syria
*Corresponding author: ghaida.rebdawi@hiast.edu.sy

Received December 31, 2012; Revised April 15, 2013; Accepted April 17, 2013

Abstract Dictionaries are very essential resources that almost all Natural Language Processing (NLP) applications use. Since language is constantly evolving, new words or new meanings to current words continuously appear. In order to keep a dictionary up-to-date, an enrichment process is needed to incorporate new vocabularies. In the last decade, a new approach of resources construction has emerged based on the collaboration between different users on the Web. In this paper, we present the Interactive Arabic Dictionary (IAD): a monolingual web-based dictionary. Initially based on the “Almuajam Alwasseet” dictionary, IAD provides the different meanings of Arabic words, with specific morphological and syntactical information, in addition to other related information such as example sentences, multimedia illustrations, associated words, semantic domains, expressions, linguistic avails, common mistakes. Authorized users can collaboratively enrich the content of the dictionary through the use of a “controlled process” to add or modify entries, meanings, or any kind of detailed information related to them. This “controlled process” consists of a suggestion-validation procedure in order to maintain the integrity of the dictionary. This enrichment process will expand the dictionary content, allowing its future exploitation in high level NLP applications.

Keywords: *collaboratively constructed language resources, Arabic dictionary, semantic lexical resource, interactive dictionary, Almuajam Alwasseet*

1. Introduction

Semantic lexical resources are very important for various Natural Language Processing (NLP) applications. However, such comprehensive and trustworthy resources are rare, and not often freely available. The cost of constructing these resources manually is very high, and building them automatically requires exhaustive validation by experts.

In the last decade, a new approach of resource construction has emerged. Resources were constructed progressively, by the interaction of users with applications on the Web. Web technologies supported distributed collaboration and made it accessible to Internet users. In fact, the overhead of conventional language resource construction can be overcome by collaboration. Wikipedia [5], the free encyclopedia, that can be edited by anyone, is the most popular and promising resource in this respect.

Designed as the lexical companion to Wikipedia, Wiktionary [6], the wiki dictionary, is a multilingual, web-based project to create a free content dictionary, available in 158 languages. Unlike standard dictionaries, it is written collaboratively by volunteers, called “Wiktionarians”, using wiki software, allowing articles to be changed by almost anyone with access to the website. Wiktionary has grown beyond a standard dictionary and now includes a thesaurus, a rhyme guide, phrase books, language statistics and extensive appendices. It is intended to include not only the definition of a word, but also enough information to really understand it, such as etymologies, pronunciations, sample quotations, synonyms, antonyms and translations. An Arabic version of Wiktionary [7] is available, but has a limited number of entries.

The Quranic Arabic Corpus [4] is another collaboratively constructed linguistic resource initiated at the University of Leeds, with multiple layers of annotation including part-of-speech tagging, morphological segmentation [8] and syntactic analysis using dependency grammar [9]. The motivation behind this work is to produce a resource that enables further analysis of the Quran.

Collaboratively constructed resources face two main challenges: (1) the integration of recently added content with the existing resource, and (2) the quality of the acquired content. Collaboratively constructed resources often lack quality or contain incomplete entries as they are often accessible to non-expert users and lack editorial control. To overcome these problems and acquire valuable knowledge, researchers and linguists in HIAST (Higher Institute for Applied Sciences and Technology) launched a project [3,15] to build an Interactive Arabic dictionary (IAD) based initially on the "Almuajam Alwasseet" [10], which can be collaboratively enriched with new entries, meanings, and other morphological, syntactical, semantic information.

2. Interactive Arabic Dictionary

Many studies were carried out to specify the main characteristics and features of the Arabic dictionary. “Constructing Computer Dictionary” [1], “Specification of the Interactive Arabic Dictionary” [11], and “Conceptual Design of the Interactive Arabic Dictionary” [12], were the main studies used in HIAST to implement the Interactive dictionary.

2.1. Objectives

IAD is a Monolingual dictionary (Arabic-Arabic), targeted to Arabic language speakers and learners. This dictionary contains a multi-level linguistic knowledge: morphological, lexical, syntactical, and semantic. Moreover, it provides many linguistic statistics useful for linguistic researchers and software developers.

IAD offers the possibility of searching word meaning, extended with a number of illustrative examples (that presents the correct use of the word in Arabic), and some multimedia contents (images, sounds, videos). It also provides other information, such as non-standard plural forms, associated words, semantic relations (synonyms, antonyms), idioms, common mistakes, linguistic tips.

IAD includes a simplified version of the morphological analyzer - developed at HIAST [13,14] - to extract the stem of the given word, and a spelling checker to check the spelling of the searched word and propose alternatives. IAD is also integrated with the open source system for derivation and conjugation "SARF" [2] to enable access to the derivation and conjugation of the searched words.

2.2. Main Actors

The Arabic Interactive Dictionary is designed to allow a real interaction with web users searching for Arabic word meanings. Users with high privileges (Linguists /Lexicographers) can also enrich the dictionary with new words, meanings, examples, multimedia, or other related information. From this perspective, it was necessary to design a system that can manage the interactivity preserving the correctness and integrity of the dictionary.

Users of the system can be categorized into 4 categories: Common users, Linguists, Lexicographers, and Administrators.

Common users access the dictionary through a web interface to search for word meanings and other related information.

Linguists can suggest insertion of new words, update of an existing word meaning, or other related information. To access the system as a linguist, the user should apply for a linguist account (providing a username, a password, and other required information).

Lexicographers can validate or reject linguists' propositions. Lexicographers can also provide configurations to derive specialized dictionaries. The administration committee of the dictionary designates users to access the system as lexicographers.

Administrators can manage the accounts of the system users.

2.3. Data Model

The dictionary was designed to be consistent with the morphological and semantic characteristics of Arabic language [3]. This design has adopted the word generation model in Arabic, and the basic rules and patterns stated in [1].

The lexical entry of the dictionary is a word that could be a verb, a noun, or a preposition. Each entry is associated with a root, a pattern, a diacritized form, and one (or more) meaning.

In case the word is a verb, other information are associated such as: present form, infinitive form, transitivity feature (zero, one, two, or three objects), and associated nouns.

In case the word is a noun, other information are associated such as: gender (masculine, feminine), number (single, pair, plural), type (Instrumental noun, Gerund) origin (Arabized, imported), and associated verbs.

As mentioned earlier, each entry has one (or more) meaning. A word coupled with a specific meaning has attributes, such as:

- gloss
- plural form
- usage frequency
- gloss reference (referenced dictionary)
- domain of use (specialization)
- etymological information
- examples of use (with corresponding references and multimedia illustrations)
- multimedia (audio for sound expressing words, video, image)
- common mistakes
- linguistic tips
- semantic domain

2.4. IAD Functions

The dictionary provides four main functions: searching, enrichment, access to the derivation and conjugation system “SARF”, and statistics.

2.4.1. Search Function

Searching function is available for all dictionary users. IAD offers two types of searching: by entry and by root.

Searching *by entry* returns all dictionary entries that match the entered string, with access to related meanings and other information. The searched string may consist of one or many words forming an expression or an idiom.

Searching *by root* returns all dictionary entries derived from the entered root with access to related meanings and other information.

Accessing one of the meanings in the returned list displays the gloss that defines the entry meaning, in addition to related examples, morphological and syntactical information, multimedia, associated words, semantic relations, expressions, idioms, linguistic tips, and common mistakes.

When the entry meaning is displayed, IAD provides an access to semantic search, which enables searching synonyms, antonyms, or other semantically related entries.

The advanced search option allows restricting the search space to either one of the categories: verbs, nouns, prepositions, or expressions.

Searching scenario can be presented by the following steps:

a) User enters an Arabic word that can be completely or partially diacritized.

b) User decides the search type (by entry or by root).

c) The system tries to match the word with the dictionary entries or roots

i) If it finds a match, a list of corresponding entries is displayed (see Figure 1):

● In case of searching by entry, this list contains all possible entries (verbs, nouns, prepositions, etc.) that correspond to the searched word letters and diacritics (if specified).

● In case of searching by root, this list contains all possible entries (verbs, nouns, prepositions, etc.) derived from this root.

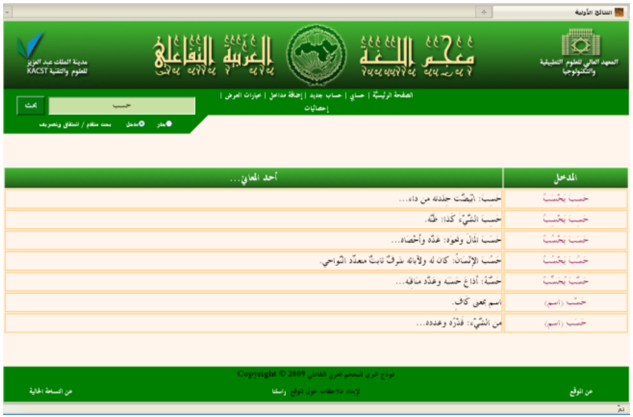


Figure 1. List of corresponding entries interface

ii) If the given word does not match any entry, the embedded morphological analyzer is called to determine the stem of the word.

● If the stem exists in the dictionary entries or roots, the system proceeds with the search using the stem.

● Otherwise, the embedded spelling checker is called to suggest alternatives. User selects the desired word and the system proceeds with a new search operation using the desired word (see Figure 2).

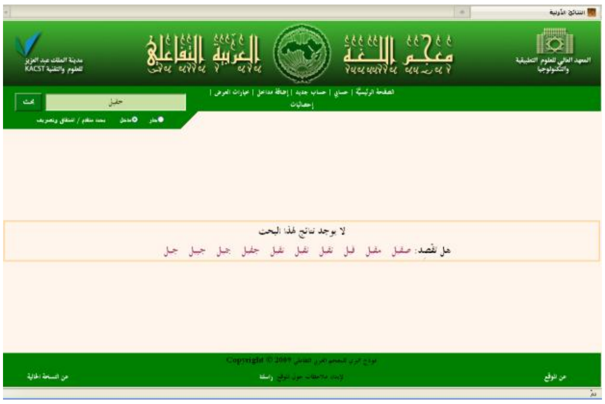


Figure 2. Alternatives suggested by the spelling checker

d) User selects the desired diacritized entry from the retuned list.

e) The system moves to a detailed information page containing the morphological characteristics and the different meanings of the word and other related information (see Figure 3).

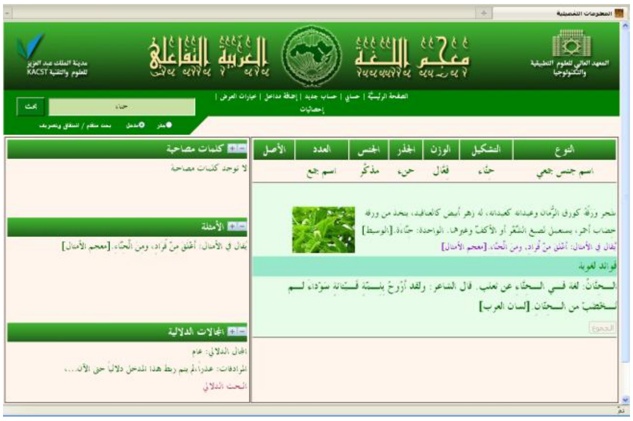


Figure 3. Detailled information page

2.4.2. Enrichment

Linguists can enrich the dictionary with new words, meanings, semantic relations, or any other information following a mechanism that ensures security, coherence, and integrity.

Only registered linguists can modify the dictionary content. Linguist can suggest adding/modifying all kinds of detailed information related to entries and meanings (see Figure 4), such as:

● Adding new entries, with corresponding meanings, examples, morphological information,

● Adding new meanings to existing words with other related information.

● Modifying current content of dictionary words.

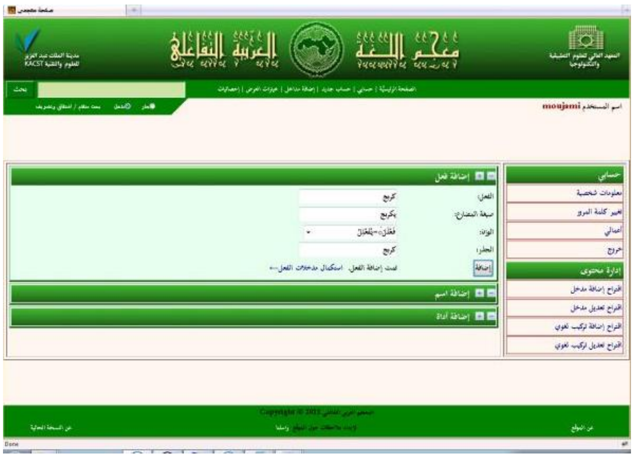


Figure 4. Enrichment suggestion interface

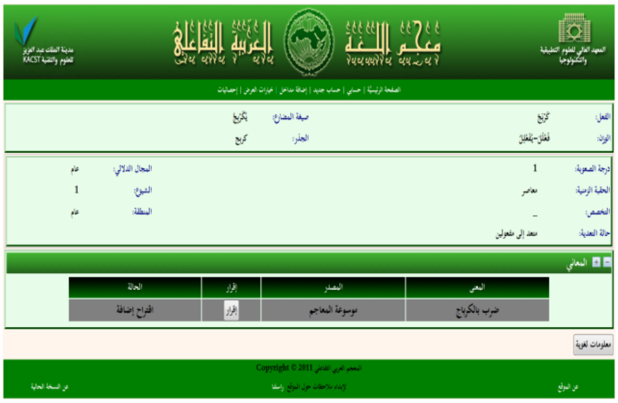


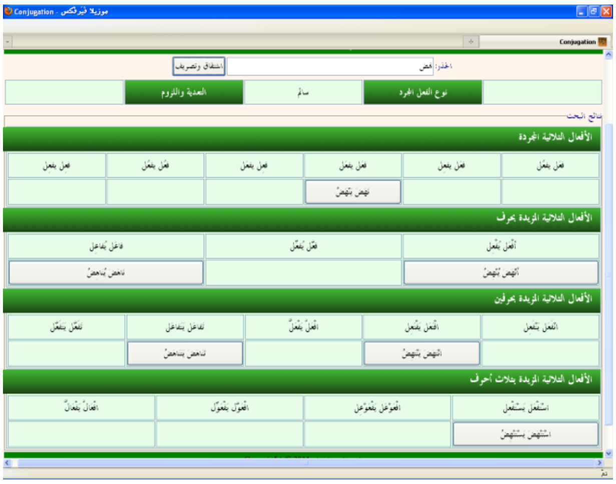
Figure 5. Enrichment validation interface

Enrichment is an interactive process, in which available related information is presented to the user in order to guide him through the suggestion process. These suggestions will be labeled as “pending”, and will not be incorporated in the database of the dictionary until the approval of the lexicographer who can explore the suggestions and then accept, modify, or reject them (see Figure 5). The approved suggestions will be part of the dictionary, and will appear in the results of next search operations.

2.4.3. Access to "SARF"

The Interactive Arabic Dictionary is integrated with the open source system for derivation and conjugation "SARF" [2] to enable access to the derivation and conjugation of the searched words.

The original version of SARF is a desktop application. New web interfaces were designed to integrate SARF with the dictionary to make it available to dictionary users (see Figure 6).



Conjugation - موزيلا فيرفكس

Conjugation

الجذر: نهض [اشتقاق وتصريف]

	التعدية واللزوم	سالم	نوع الفعل المجرد	

نتائج البحث

الأفعال الثلاثية المجردة

فَعِلَ يَفْعِلُ	فَعُلَ يَفْعُلُ	فَعِلَ يَفْعَلُ	فَعَلَ يَفْعَلُ	فَعَلَ يَفْعِلُ	فَعَلَ يَفْعُلُ
			نَهَضَ يَنْهَضُ		

الأفعال الثلاثية المزيدة بحرف

فاعَلَ يُفاعِلُ	فَعَّلَ يُفَعِّلُ	أَفْعَلَ يُفْعِلُ
ناهَضَ يُناهِضُ		أَنْهَضَ يُنْهِضُ

الأفعال الثلاثية المزيدة بحرفين

تَفَعَّلَ يَتَفَعَّلُ	تَفاعَلَ يَتَفاعَلُ	افْعَلَّ يَفْعَلُّ	افْتَعَلَ يَفْتَعِلُ	انْفَعَلَ يَنْفَعِلُ
	تَناهَضَ يَتَناهَضُ		انْتَهَضَ يَنْتَهِضُ	

الأفعال الثلاثية المزيدة بثلاث أحرف

افْعالَّ يَفْعالُّ	افْعَوَّلَ يَفْعَوِّلُ	افْعَوْعَلَ يَفْعَوْعِلُ	اسْتَفْعَلَ يَسْتَفْعِلُ
			اسْتَنْهَضَ يَسْتَنْهِضُ

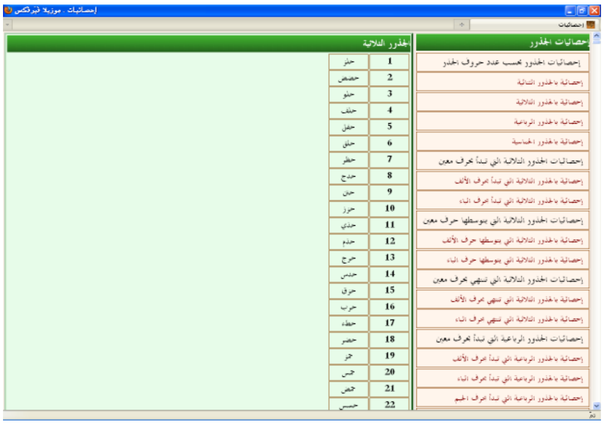
تم

Figure 6. SARF web interface

2.4.4. Statistics

The Interactive Arabic Dictionary allow users to acquire statistical information about its content (see Figure 7). The dictionary provides statistics related to different aspects, such as: root number of letters (Tri-literal, Quadr-literal, Quinque-literal), types of verbs (augmented, or unaugmented), location of characters in the root.

Another category of statistics is provided in the dictionary to quantify the contribution of linguists and lexicographers in the enrichment process.



إحصائيات - موزيلا فيرفكس

إحصائيات

إحصائيات الجذور

- إحصائيات الجذور بحسب عدد حروف الجذر
- إحصائية بالجذور الثنائية
- إحصائية بالجذور الثلاثية
- إحصائية بالجذور الرباعية
- إحصائية بالجذور الخماسية
- إحصائيات الجذور الثلاثية التي تبدأ بحرف معين
- إحصائية بالجذور الثلاثية التي تبدأ بحرف الألف
- إحصائية بالجذور الثلاثية التي تبدأ بحرف الباء
- إحصائيات الجذور الثلاثية التي يتوسطها حرف معين
- إحصائية بالجذور الثلاثية التي يتوسطها حرف الألف
- إحصائية بالجذور الثلاثية التي يتوسطها حرف الباء
- إحصائيات الجذور الثلاثية التي تنتهي بحرف معين
- إحصائية بالجذور الثلاثية التي تنتهي بحرف الألف
- إحصائية بالجذور الثلاثية التي تنتهي بحرف الباء
- إحصائيات الجذور الرباعية التي تبدأ بحرف معين
- إحصائية بالجذور الرباعية التي تبدأ بحرف الألف
- إحصائية بالجذور الرباعية التي تبدأ بحرف الباء
- إحصائية بالجذور الرباعية التي تبدأ بحرف الجيم

الجذور الثلاثية

حنز	1
حضض	2
حلو	3
حلف	4
حفل	5
حلق	6
حظر	7
حدج	8
حن	9
حزز	10
حذي	11
حذم	12
حرج	13
حبس	14
حرق	15
حرب	16
حطء	17
حضر	18
حز	19
حس	20
حض	21
حسس	22

تم

Figure 7. Statistics interface

2.5. IAD Design

The system is decomposed into four subsystems: (1) search subsystem, which comprises the morphological analyzer and the spelling checker components, (2) suggestion subsystem, (3) validation subsystem, and (4) accounts' management subsystem.

These four subsystems interact with the database which includes linguistic data, data about the entries state (approved, disapproved, pending), and data about users' accounts. Figure 8 shows a diagram of system decomposition and interaction between the different subsystems.

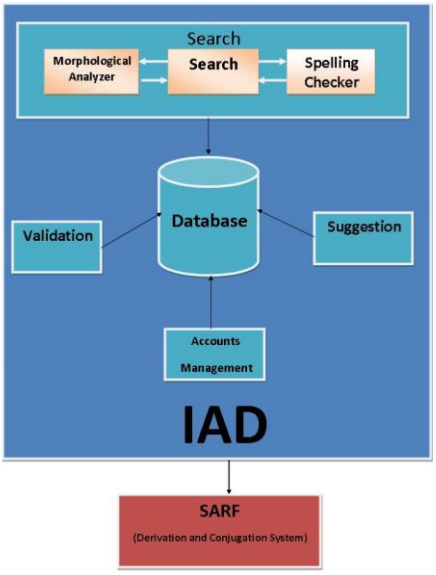


Figure 8. IAD architecture

2.6. Implementation Issues

The system is implemented using n-tiers architecture; all subsystems are divided into four tiers: Persistence Layer, Data Access Layer, Business Logic Layer, and Presentation Layer.

1. Persistence Layer: This tier is implemented using Hibernate technology which provides several facilities for data retrieval, data updating, and transaction management. Hibernate enables to generate Java source files to match the structure of the dictionary database based on object-relational mapping specified in its XML configuration files.

2. Data Access Layer (DAL): Using persistence layer, it is easy to implement data access layer. A generic class is provided to perform common tasks in data access layer. All other classes in the DAL inherit from this class and add special behavior if any. Moreover, a DAL factory, which represents a single interface between DAL and higher layers, is provided.

3. Business Logic Layer (BLL): Due to the complexity of this layer, we divide it into two sub-layers:

● BOManager layer responsible of managing business objects (create, retrieve, update, delete), and filling them with the related values from different Data objects.

● Service layer integrates between BOManagers to provide several services in the system such as search, morphological analysis, spelling correction.

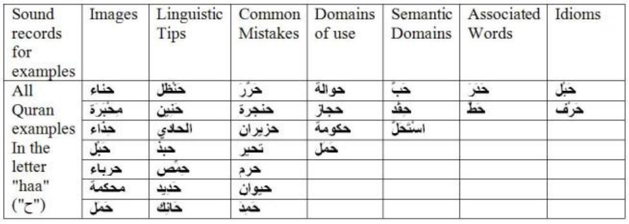
As in DAL, a single interface between BLL and higher layers is provided.

4. Presentation Layer: This layer represents the front-end of the software application. It consists of several jsp and html pages.

2.7. Dictionary Content

The current version of the Arabic Interactive Dictionary published on the site http://almuajam.hiast.edu.sy/ contains all “Almuajam Alwasseet” dictionary entries [10] enriched from other important traditional and contemporary Arabic dictionaries. Based on a paper version of “Almuajam Alwasseet”, a data engineering procedure was carried on to structure the paper dictionary content. This procedure yielded the kernel of IAD database. There are more than 50000verbs and 75000 nouns. This kernel was extended with examples extracted from many sources (Quran, Hadith, traditional and contemporary Arabic books). A special effort has been done to enrich the entries of the letter “Haa'/ح” in order to illustrate all IAD features and characteristics. Thus, many examples and multimedia illustrations were added, semantic domains for many entries were specified, and sound records for all Quran examples were provided. Table 1 presents a list of examples illustrating some of the dictionary features. The content of the dictionary is always subject to enrichment respecting integrity and correctness constraints mentioned earlier.

Table 1. A List of Examples Illustrating the Dictionary Features



Sound records for examples	Images	Linguistic Tips	Common Mistakes	Domains of use	Semantic Domains	Associated Words	Idioms
All Quran examples In the letter "haa" ("ح")	حناء	حنْظل	حرَّر	حوالة	حبّ	حذَر	حبَل
	مِحْبَرة	حنين	حنجرة	حجاز	حقد	حط	حرْف
	حذاء	الحادي	حزيران	حكومة	استَحلّ		
	حبَل	حبذ	تحير	حمَل			
	حرباء	حمّص	حرم				
	محكمة	حديد	حيوان				
	حمَل	حائك	حمد				

3. Conclusion

A Beta version of the dictionary is now available on the web site http://almuajam.hiast.edu.sy/. The dictionary is now ready to be enriched collaboratively by web users. Maintaining the integrity and correctness of the dictionary content, requires the supervision of an administration committee responsible of assigning lexicographers and specifying rules for linguists’ admission.

Research is undertaken actually at HIAST to enhance IAD performance and extend its content. A more efficient version of the morphological analyzer will be integrated with IAD. Projects are envisaged to support the enrichment process by automated tools. Using available corpora on the web, these tools enable the dictionary to be enriched with examples, other meanings, media.

To enable Arabic language processing applications to access the different functionalities of IAD, an application programming interface (API) will be provided.

Acknowledgement

The present work has been sponsored by King Abdulaziz City for Science and Technology (KACST) in Kingdom of Saudi Arabia, and Arab League Educational, Cultural and Scientific Organization (ALECSO), and HIAST (Higher Institute for Applied Sciences and Technology) in Syria.

The work is a success story of collaboration between researchers, engineers, and linguistic experts in HIAST. The authors wish to thank: M. Al-Bawab, R. Sonbol, S. Alattar, F. Al-Hassan, W. Al-Hassan, I. Waynakh, O. Rajab for their valuable efforts in the project.

References

[1] Al-Bawab M., Constructing Computer Dictionary, 2008.

[2] Derivation and conjugation system "SARF", http://sourceforge.net/projects/sarf/.

[3] Gh. Rebdawi, N. Ghneim, M. S. Desouki, R. Sonbol, S. Alattar, F. AlHassan, W. AlHassan, I. Waynakh, M. Al-Bawab, O. Rajab, Interactive Arabic Dictionary – Technical report, Internal report, HIAST, Damascus, 2011.

[4] http://corpus.quran.com.

[5] http://www.wikipedia.org.

[6] http://www.wiktionary.org.

[7] http://ar.wiktionary.org.

[8] Kais Dukes and Nizar Habash, “Morphological Annotation of Quranic Arabic,” in The Language Resources and Evaluation Conference (LREC 2010), Malta, 2010.

[9] K. Dukes and T. Buckwalter, “A Dependency Treebank of the Quran using Traditional Arabic Grammar,” in The 7th International Conference on Informatics and Systems (INFOS) 2010. Cairo, Egypt.

[10] Mustafa I., Alzayat A. H., Abdel-kader h., Alnajar M. A, Al-Wasseet dictionary, 3rd edition, Alnouri Press, Damascus, 1960.

[11] Interactive Arabic Dictionary (Project Specifications), 2008.

[12] Interactive Arabic Dictionary (Project Conceptual Design), 2009.

[13] Sonbol. R, Ghneim, N. and Desouki, M.S, “Arabic Morphological Analysis: a New Approach,” in 3d International Conference on Information and Communication Technologies: from Theory to Applications - ICTTA'08. Damascus, Syria, 2008.

[14] Sonbol. R, Ghneim, N. and Desouki, M.S., “An Application Oriented Arabic Morphological Analyzer,” Damascus University Journal, Vol. (27) No.(1), Pages 7-19, January 2011.

[15] Gh. Rebdawi, N. Ghneim, M., Desouki, R. Sonbol, An Interactive Arabic Dictionary, 7th International Conference on Innovations in Information Technology, Arabic Language Processing special session, 25-27 April, Abu Dhabi, United Arab Emirate, 2011.