Arabic WordNet: Current State and Future Extensions

Horacio Rodríguez[1], David Farwell[1], Javi Farreres[1], Manuel Bertran[1], Musa Alkhalifa[2], M. Antonia Martí[2], William Black[3], Sabri Elkateb[3], James Kirk[3], Adam Pease[4], Piek Vossen[5], and Christiane Fellbaum[6]

[1]Politechnical University of Catalonia
Jordi Girona, 1-3; 08034 Barcelona; Spain
{horacio, farwell, farreres, mbertran@lsi.upc.edu}

[2]Universita de Barcelona
Despatx: 5.19 Edifici Josep Carner, Gran Via 585; 08007 Barcelona; Spain
musa@thera-clic.com, amarti@ub.edu

[3]The University of Manchester
PO Box 88, Sackville St; Manchester, M60 1QD; UK
{w.black, sabri.elkateb, James.E.Kirk@manchester.ac.uk}

[4]Articulate Software Inc,
420 College Ave; Angwin, CA 94508; USA
apease@articulatesoftware.com

[5]Irion Technologies
Delftechpark 26; 2628XH, Delft, The Netherlands
piek.vossen@irion.nl

[6]Princeton University,
Department of Psychology, Green Hall; Princeton, NJ 08544; USA
fellbaum@clarity.princeton.edu

Abstract. We report on the current status of the Arabic WordNet project and in particular on the contents of the database, the lexicographer and user interfaces, the Arabic WordNet browser, linking to the SUMO ontology, the Arabic word spotter, and techniques for semi-automatically extending Arabic WordNet. The central focus of the presentation is on the semi-automatic extension of Arabic WordNet using lexical and morphological rules.

Keywords: Arabic NLP, Arabic WordNet, Ontology, Semi-automatic WordNet extension.

1. Introduction

Arabic WordNet (AWN – [1], [2], [3], inter alia) is currently under construction following a methodology developed for EuroWordNet [4]. The EuroWordNet approach maximizes compatibility across wordnets and focuses on the manual encoding of a set of base concepts, the most salient and important concepts as defined by various network-based and corpus-based criteria as reported in Rodríguez, et al [5]. Like EuroWordNet, there is a straightforward mapping from Arabic WordNet (AWN) onto Princeton WordNet 2.0 (PWN – [6]). In addition to constructing a WordNet for Arabic, the AWN project aims to extend a formal specification of the senses of its synsets using the Suggested Upper Merged Ontology (SUMO), a language-independent ontology. This representation is essentially an interlingua between all wordnets ([7], [8]) and can serve as the basis for developing semantics-based computational tools for cross-linguistic NLP applications[1].

The following discussion is divided into two main parts. We first present the current status of the Arabic WordNet and then we describe different techniques for semi-automatically extending AWN.

2. Current State of Arabic WordNet

2.1 Content of the Arabic WordNet Database

At the time of writing Arabic WordNet consists of 9228 synsets (6252 nominal, 2260 verbal, 606 adjectival, and 106 adverbial), containing 18,957 Arabic expressions. This number includes 1155 synsets that correspond to Named Entities which have been extracted automatically and are being checked by the lexicographers. Since these numbers are constantly changing, the interested reader can find the most up-to-date statistics at: http://www.lsi.upc.edu/~mbertran/arabic/awn/query/sug_statistics.php.

2.2 Interfaces

Two different web-based interfaces have been developed for the AWN project.

Lexicographer's Web Interface (Barcelona)
http://www.lsi.upc.edu/~mbertran/arabic/awn/update/synset_browse.php

[1] To our knowledge the only previous attempt to build a wordnet for the Arabic language consisted of a set of experiments by Mona Diab, [9] for attaching Arabic words to English synsets using only English WordNet and a parallel corpus Arabic English as knowledge sources.

The lexicographer's interface has been designed to support the task of adding, modifying, moving or deleting WordNet synsets. Its functionalities include:

- listing the synsets assigned to each lexicographer (here, the lexicographer has many options to select from, including listing 'completed synsets' or 'incomplete synsets' or both),
- listing synsets by English word,
- listing synsets by synset offsets,
- listing synsets by date of creation,
- listing synsets without associated lexical items, or yet to be reviewed (to enhance validation, each lexicographer can review and comment on the others' entries).

User's Web Interface (Barcelona)
http://www.lsi.upc.edu/~mbertran/arabic/awn/index.html
This interface enables the user to consult AWN and search for Arabic words, Arabic roots, Arabic synsets, English words, synset offsets for English WordNet 2.0. Search can be refined by selecting the appropriate part of speech. A virtual keyboard is also available for users who do not have access to an Arabic keyboard.

2.3 WordNet to SUMO Mapping

SUMO ([7], [10]) and its domain ontologies form the largest publicly available formal ontology today. It is formally defined and not dependent on a particular application. SUMO contains 1000 terms, 4000 axioms, 750 rules and is the only formal ontology that has been mapped by hand to all of the PWN synsets as well as to EuroWordNet and BalkaNet. However, because WordNet is much larger than SUMO, many links are from general SUMO terms to more specific WordNet synsets. As of this writing, there are 3772 equivalence mappings, 100,477 subsuming mappings, and 10,930 mappings from a SUMO class to a WordNet instance. Most nouns map to SUMO classes, most verbs to subclasses of processes, most adjectives to subjective assessment attributes, and most adverbs to relations of and manners. While instance mappings are often from very specific SUMO classes, SUMO itself only includes a few sets of instances, such as the countries of the world. SUMO and its associated domain ontologies have a total of roughly 20,000 terms and 70,000 axioms.

SUMO synset definitions of the relevant synset can be viewed from the user's web interface by using the SUMO Search Tool which relates PWN synsets to concepts in the SUMO ontology. To facilitate understanding of the ontology by Arabic speakers, the Sigma ontology management system [10] automatically generates Arabic paraphrases of its formal, logical axioms. SUMO has been extended with a number of concepts that correspond to words lexicalized in Arabic but not in English. They include concepts related to Arabic/Muslim cultural and religious practices and kinship relations. This is one way in which having a formal ontology provides an interlingua that is not limited by the lexicalization of any particular human language. For more information, see:

http://sigmakee.cvs.sourceforge.net/*checkout*/sigmakee/KBs/ArabicCulture.kif

2.4 The AWN Browser

The Arabic WordNet Browser is a stand-alone application that can be run on any computer that has a Java virtual machine. In its current state, its main facilities include browsing AWN, searching for concepts in AWN, and updating AWN with latest data from the lexicographers.

Searching can be done using either English or Arabic. In Arabic, the search can be carried out using either Arabic script or Buckwalter transliteration [11] and can be for a word or root form, with the optional use of diacritics. For English, the browser supports a word-sense search alongside a graphical tree representation of PWN which allows a user to navigate via hyponym and hypernym relations between synsets. A combination of word-sense search and tree navigation enables a user to quickly and efficiently browse translations for English into Arabic.

Since users unfamiliar with Arabic cannot be expected to know how to convert an Arabic word they have copied from a Web page into an appropriate citation form, we have integrated Arabic morphological analysis into the search function, using a version of AraMorph [12]. A virtual Arabic keyboard is also accessible to enable Arabic script entry for the different search fields.

SUMO ontology navigation is currently being integrated into the browser, using a tree traversal procedure similar to that for PWN. Users will be able to search or browse AWN using SUMO as the interlingual index between English and Arabic. Also under construction are Arabic tree navigation and the automatic generation of Arabic glosses. These additions will be included in the next release version of the browser.

More detailed information and screen shots can be found at:
http://www.globalwordnet.org/AWN/AWNBrowser.html
The browser is available for downloading from Sourceforge under the General Public License (GPL) at: http://sourceforge.net/projects/awnbrowser/

2.5 The Arabic Word Spotter

An Arabic Word Spotter has been developed to provide the user with a tool to test AWN's coverage by identifying those words in an Arabic web page that can be found in AWN. The word spotter can be accessed at:
http://www.lsi.upc.edu/~mbertran/arabic/wwwWn7/

Arabic words are searched for first in AWN and, failing that, in a few bilingual dictionaries. The procedure relies on the AraMorph stemmer and, once a match is found, a word level translation is provided. Translation of stop words is provided as well.

Help and HowTos are available from:
http://www.lsi.upc.edu/~mbertran/arabic/wwwWn7/help/help.php?

3 Approaches to the Semi-automatic Extension of AWN

Although the construction of AWN has been manual, some efforts have been made to automate part of the process using available bilingual lexical resources. Using lexical resources for the semi-automatic building of wordnets for languages other than English is not new. In some cases a substantial part of the work has been performed automatically, using PWN as source ontology and bilingual resources for proposing correlates. An early effort along these lines was carried out during the development of Spanish WordNet within the framework of EuroWordNet project ([13], [5]). Later, the Catalan WordNet [14] and Basque WordNet [15] were developed following the same approach.

Within the BalkaNet project [16] and the Hungarian WordNet project [17], this same methodology was followed. In this case, the basic approach was complemented by methods that relied on monolingual dictionaries. As an experiment with the Romanian WordNet, [18] follow a similar approach, but use additional knowledge sources including Magnini's WordNet domains [19] and WordNet glosses. They use a set of metarules for combining the results of the individual heuristics and achieve 91% accuracy for the 9610 synsets covered. Finally, to build both a Chinese WordNet and a Chinese-English WordNet, [20] complement their bilingual resources with information extracted from a monolingual Chinese dictionary.

For AWN, we have investigated two different possible approaches. On the hand, we produce lists of suggested Arabic translations for the different words contained in the English synsets corresponding to the set of Base Concepts. In this case the input to the lexicographical task is the English synset, its set of synonyms and their Arabic translations. On the other hand, we derive new Arabic word forms from already existing, manually built, Arabic verbal synsets using inflectional and derivational rules and produce a list of suggested English synset associations for each form. In this case the input is the Arabic verb, the set of possible derivates and the set of English synsets which would be linked to corresponding Arabic synset. In both cases, the list of suggestions is manually validated by lexicographers.

3.1 Suggested Translations

For this approach, we start with a list of <English word, Arabic word, POS> tuples extracted from several publicly available English/Arabic resources. The first step was to clean and standardize the entries. The available resources differ in many details. Some contain POS for each entry while others do not. Arabic words were in some cases vocalized and in others not. In some cases certain diacritics are used, such as shadda (i.e., consonant reduplication), while in others no diacritics at all appear. Some dictionaries contain the perfect tense form for verbs while others use the imperfect form. After this standardization process, we merged all the sources (using both directions of translation) into one single bilingual lexicon and then took the intersection of this lexicon with the set of Base Concept word forms This latter set

was built merging the Base Concepts of EuroWordNet, 1024 synsets, with those of Balkanet, 8516 synsets.

Following 8 heuristic procedures used in building the Spanish WordNet [21] as part of EuroWordNet [4], the associations between Arabic words and PWN synsets in the Arabic-English bilingual lexicon were scored. The methodology assigned a score to each association, but since the Arabic WordNet has been manually constructed, no threshold was set and all associations were provided to the lexicographer for verification. Thus, when editing an Arabic synset, the lexicographer begins with a suggested association, rather than an empty synset with only the English data to go by. Some suggestions were correct or very similar to correct ones. Others were incorrect but served to trigger an Arabic word that might otherwise have been missed. The result has been a much richer set of Arabic synsets.

Initially 15,115 translations were suggested, of which only 9748 (64.5%) have been thus far checked by the lexicographers. The results show that of these, 392 candidates (4.0%) were accepted without any changes, 1246 (12.8%) were accepted with minor changes (such as adding diacritics), 877 (9.0%), while good candidates, were rejected because they were identical or very similar to translations that had already been chosen by the lexicographer, and 7233 (74.2%) were rejected because they were incorrect given the gloss and examples. We will revise these results once all the Base concepts have been completed at the end of the project.

At first glance, these results are not especially impressive and, as a result, we turned to an alternative approach. At the same time, it is difficult to compare these figures with results obtained for other languages because we are interested exclusively in generating suggestions for Base Concepts which are to be confirmed by lexicographers while other approaches do not have this objective. Since the words belonging to Base Concept synsets are often highly polysemous, the accuracy of predicting translations is generally lower. In addition, since we are more interested in high coverage, no filters were applied with a corresponding drop in precision.

3.2 Semi-automatic Extension of AWN Using Lexical and Morphological Rules

In this section we explore an alternative methodology for the semi-automatic extension of Arabic WordNet using lexical rules as applied to existing AWN entries. This methodology takes advantage of one of a central characteristic of Arabic, namely that many words having a common root (i.e. a sequence of typically three consonants) have related meanings and can be derived from a base verbal form by means of a reduced set of lexical rules. Since AWN entries must be manually reviewed, our aim is once again not to automatically attach new synsets but rather to suggest new attachments and to evaluate whether these suggestions can help the lexicographer. As with previous approach, we are more interested in getting a broad coverage than high accuracy, although an appropriate balance between these two measures is nonetheless desirable.

3.2.1 Setting

In the studies reported in this section, we deal only with a very limited but highly productive set of lexical rules which produce regular verbal derivative forms, regular nominal and adjectival derivative forms and, of course, inflected verbal forms.

From most of the basic Arabic triliteral verbal entries, up to 9 additional verbal forms can be regularly derived as shown in Table 1. We refer to the set of lexical rules that account for these forms as Rule Set 1. They have been implemented as regular expression patterns.

Class	Arabic Pattern
1 (Basic)	فعل
2	فعّل
3	فاعل
4	افعل
5	تفعّل
6	تفاعل
7	انفعل
8	افتعل
9	افعلّ
10	استفعل

Table 1: Patterns of Arabic regular derived forms

For instance, the basic form درس (DaRaSa, to study/to learn) has as its root درس (DRS). The first form pattern in Table 1 applied to this root produces the original basic forms (in this case simply adding diacritics). If we apply the second form pattern in Table 2 to the same root, the form درّس (DaRRaSa, to teach) is obtained.

From any verbal form (whether basic or derived by Rule Set 1), both nominal and adjectival forms can also be generated in a highly systematic way: the nominal verb (masdar) as well as masculine and feminine active and passive participles. We refer to this set of rules as Rule Set 2. Examples include the masdar درس (DaRSun, lesson, study) from درس (DaRaSa, to study/to learn) and مدرّس (MuDaRRiSun, male teacher) from درّس (DaRRaSa, to teach).

Finally, a set of morphological rules for each basic or derived verb form is applied in order to produce the full set of inflected verb forms as exemplified in Table 2.

Table 2: Some inflected verbal forms (of 82 possible) for درس (DaRaSa, to learn)

English form	Arabic form
(he) learned	د ر س
(I) learned	د ر ست
(I) learn	ا د ر س
(he) learns	يـد ر س
(we) learn	نـد ر س
...	...

As reported below, these forms are especially useful for searching a corpus as well as in various applications. The number of different forms depends on the class of the verb but it ranges from 44 to 84 forms. Class 1, for instance, has 82 forms and, thus, requires the application of 82 different morphological rules. We refer to this set of rules as Rule Set 3.

Beyond this, we aim to extend this basic approach to the derivation of additional forms including the feminine form from any nominal masculine form (for instance, مدرّسة, MuDaRRiSatun, female teacher, from مدرّس, MuDaRRiSun, male teacher), or the regular plural forms from any nominal singular form. For instance, the regular nominative plural form is created by adding the suffix (Una) to the singular form (e.g., مدرّسون MuDaRRiSUna, male teachers, is derived from مدرّس, MuDaRRiSun, male teacher).

3.2.2 Central Problems to Address

Implementing the ideas stated in the previous section is not straightforward. Several problems have to be addressed but perhaps the two most important are 1) filtering noise caused by over the generation of derivative verb forms and 2) mapping the newly created Arabic word forms to appropriate WordNet synsets, i.e., mapping words to their appropriate sense. Obviously not all the derivative forms generated by Rule Sets 1 and 2 are valid for any given basic verbal form in Arabic. For instance, for درس (DaRaSa, to learn) of the nine possible derivates generated by the application of Rule Set 1, shown in Table 1, only the six shown in Table 3 are valid according to [22]. Thus, some kind of filtering has to take place in order to reduce the noise wherever possible. That is to say, only the most promising candidates should be proposed to the lexicographer. In addition, once the set of candidate derivates has been built and the corresponding nominal and adjectival forms generated, we have to map all these forms to English translations and from these to the appropriate PWN synsets.

Table 3: Valid derivates from درس (DaRaSa, to learn)

Class	English form	Arabic form
1 (basic)	to learn, to study	د ر س
2	to teach	د رَّ س
3	to study (together with	د ا ر س

	someone)	
4	to learn with	ا د ر س
6	to study (carefully together)	تـد ا رّ س
7	to vanish	انـد ر س

3.2.3 Resources

The procedures described below make use of the following resources:

- Princeton's English WordNet 2.0,
- Arabic WordNet (specifically the set of Arabic verbal synsets currently available),
- the LOGOS database of Arabic verbs which contains 944 fully conjugated Arabic verb (available at: http://www.logosconjugator.org/verbi_utf8/all_verbs_index_ar.html),
- the NMSU bilingual Arabic-English lexicon (available at: http://crl.nmsu.edu/Resources/dictionaries/download.php?lang=Arabic),
- the Arabic GigaWord Corpus (available through LDC: http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC2006T02).

3.2.4 Overview of the Approach

Broadly speaking, the procedure we follow in generating a set of likely <Arabic word, English synset> pairs is to:

1. produce an initial list of candidate word forms (as described in Section 3.2.6),
2. filter the less likely candidates from this list (as described in Section 3.2.7),
3. generate an initial list of candidate synsets attachments (as described in Section 3.2.8),
4. score the reliability of these candidates (as described in Section 3.2.9),
5. manually review the candidates and include the valid associations in AWN.

3.2.4.1 Building the initial set of word candidates

To build the initial set of candidate word forms, we first collect a set of basic (Class 1) verb forms, such as درس (DaRaSa, to learn), from the existing 2296 verbs in AWN and transliterate them using Buckwalter encoding [11]. We next apply Rule Set 1 to generate the 9 basic derivative verb forms (both valid or not). Then, for each of these new verb forms, we apply Rule Set 3 in order to derive the full set of possible inflected forms.

3.2.4.2 Learning filters on translations

In order to determine whether or not a particular possible word form is likely to turn out to be a valid word form, we build a decision tree classifier using machine learning for each of the 9 classes of derivation (i.e. Classes 2 through 10). The choice of decision trees is mainly motivated because their ease of interpretation, since otherwise they provided similar results to those of Adaboost, an alternative approach which we have tested. We used the C5.0 implementation within Weka toolbox [23]. The software can be obtained from: http://www.cs.waikato.ac.nz/~ml/weka/index.html.

The features used for learning included the following:

1. the relative frequency of each inflected form for a given class of derivatives in the GigaWord Corpus,
2. whether the base form appears in the NMSU dictionary or not,
3. the POS tag of the base form in NMSU dictionary,
4. the attribute TRUE (positive example) or FALSE (negative example).

In order to learn a decision tree, the algorithm must be presented with both positive and negative examples. For positive examples, we used the LOGOS database (946 examples), AWN (2296 examples) and the NMSU dictionary (15,654 examples). LOGOS and AWN are the most accurate but do not provide enough material. NMSU has broad coverage but is less accurate because the entries are not vocalized and lack diacritics (for some classes the lack of the shadda diacritic[2] is a serious problem).

To build the training set, we matched each inflected form for each of the base forms (basic or derived) against the GigaWord Corpus and the NMSU dictionary in order to extract the relevant features for learning. Finally, we selected all the base forms corresponding to the word forms that occurred in the resources as positive examples, and used the remaining forms (i.e., those that do not occur in either the GigaWord Corpus or in the NMSU dictionary) as negative examples. All other forms are discarded. Table 4, for instance, shows the size of the training set used for

learning the filter for Class 7.

Table 4: Size of training set for learning the Class 7 filter

	Logos	AWN	NMSU	Total
positive	8	24	1718	1750
negative	70	0	4856	4926
Total	78	24	6574	6676

Following this general procedure, a decision tree classifier was learned for each class of derivation (in fact, only 8 filters were learned because there were too few examples for Class 9). We applied 10-fold cross-validation. The results for all the classifiers but one were over 99% of F1 value although in some cases the resultant decision tree consisted of only a single query on the occurrence of the base form in the NMSU dictionary (i.e. the form was accepted simply if it occurs in NMSU dictionary).

2 In Arabic shadda is how consonant reduplication or germination is marked. Obviously if this diacritic is lost the correct orthographic form of a word is affected

3.2.4.3 Building the list of candidate synsets attachments

To build a list of candidate synset attachments, we first generate a list of possible base verb forms by applying the filters described above. We then apply Rule Set 2 to each of the base verb forms to generate the set of related Arabic noun and adjective forms. Only those forms occurring in NMSU dictionary with English equivalents occurring in PWN are retained. For each of these word forms, all the English translations from NMSU dictionary and all their PWN synsets are collected as candidates. The result of this process is a candidate set of tuples of the form <Arabic word, English word, English synset>. The final step is to assign a reliability score to each <Arabic word, English synset> tuple.

3.2.4.4 Scoring the candidate synset attachments

Our scoring routine is based on the observation that in most cases the set of derivative forms have semantically related senses. For instance, درس (DaRaSa, to study) belongs to Class 1 and its masdar is درس (DaRSun, lesson). درّس (DaRRaSa, to teach) belongs to Class 2 and its masculine active participle is مدرّس (MuDaRRiSun, male teacher). Clearly these four words are semantically related. Therefore, if we map Arabic words to English translations and then to the corresponding PWN synsets, we can expect that the correct assignments will correspond to most semantically related synsets. In other words, the most likely <Arabic word, English synset> associations are those corresponding to the most semantically related items.

There are three levels of connections to be considered[3]:

- relations between an Arabic word and its English translations,
- relations between an English word and its PWN synsets,
- relations between a PWN synset and other synsets in PWN.

To identify the "most semantically related" associations between Arabic words and PWN synsets, we:

1. collect the set of <Arabic word, English word, English synset> tuples for a given Arabic base verb form and its derivatives,
2. extract the set of English synsets and identify all the existing semantic relations between these synsets in PWN[4],
3. build a graph with three levels of nodes corresponding to Arabic words, English words, and English synsets respectively and edges corresponding to the translation relation between Arabic words and English words, the membership relation between English words and PWN synsets and finally, the recovered relations between PWN synsets.

These are represented in the graph in Figure 1.

---

3 The relations A_{base} -> A_i have not been considered explicitly because A_{base} comes from an existing AWN synset and thus its association has already been established manually.

4 As in the rest of experiments reported in this paper we have used the relations present in PWN2.0

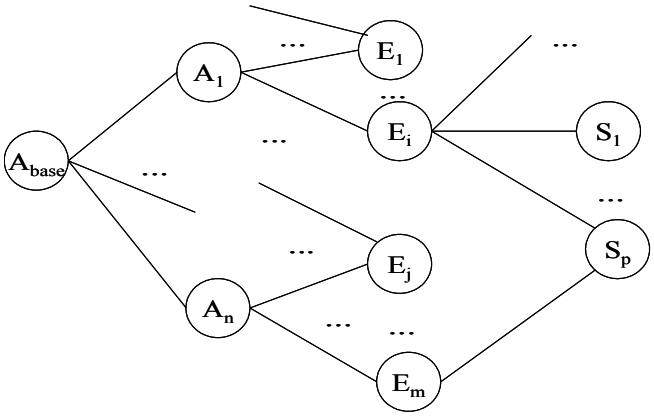

Fig. 1. Example of Graph of dependencies

Two approaches to scoring are being examined. The first, described below, is based on a set of heuristics that use the graph structure directly while the second, more complex, maps the graph onto a Bayesian Network and applies a learning algorithm. The latter approach is the subject of ongoing research and will be described in a separate forthcoming paper.

Using the graph as input, the first approach to calculating the reliability of association between Arabic word and PWN synset consists of simply applying a set of five graph traversal heuristics. The heuristics are as follows (note that in what follows, "A_{base}", "A_1", "A_2", etc., correspond to Arabic word forms, A_{base} being the initial verbal base form, "E", "E_1", "E_2", etc. to English word forms, and "S", "S_1", "S_2", etc. to PWN synsets):

1. If a unique path A-E-S exists (i.e., A is only translated as E), and E is monosemous (i.e., it is associated with a single synset), then the output tuple <A, S> is tagged as 1. See Figure 2.

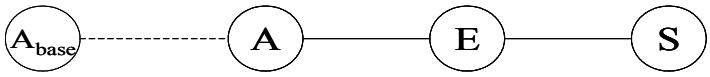

Fig. 2. Graph for heuristic 1

2. If multiple paths A-E_1-S and A-E_2-S exist (i.e., A is translated as E_1 or E_2 and both E_1 and E_2 are associated with S among other possible associations) then the output tuple <A, S> is tagged as 2. See Figure 3.

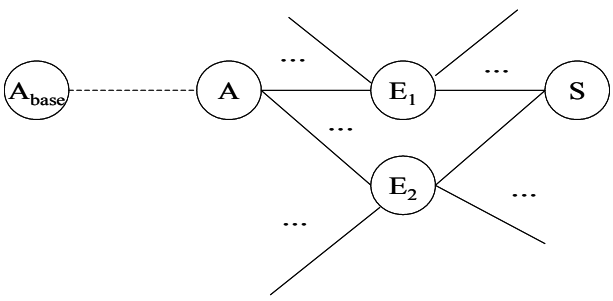

Fig. 3. Graph for heuristic 2

3. If S in A-E-S has a semantic relation to one or more synsets, S_1, S_2 … that have already been associated with an Arabic word on the basis of either heuristic 1 or heuristic 2, then the output tuple <A, S> is tagged as 3. See Figure 4.

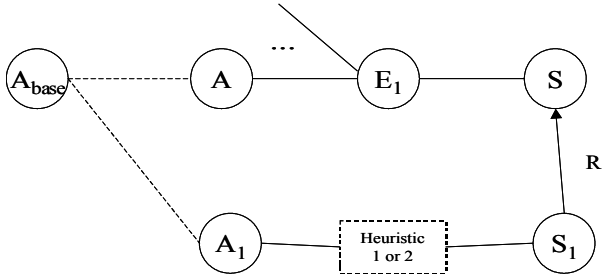

Fig. 4. Graph for heuristic 3

4. If S in A-E-S has some semantic relation with S_1, S_2 … where S_1, S_2 … belong to the set of synsets that have already been associated with related Arabic words, then the output tuple <A-S> is tagged as 4. In this case there is only one translation E of A but more than one synset associated with E. This heuristic can be sub-classified by the number of input edges or supporting semantic relations (1, 2, 3, ...). See Figure 5.

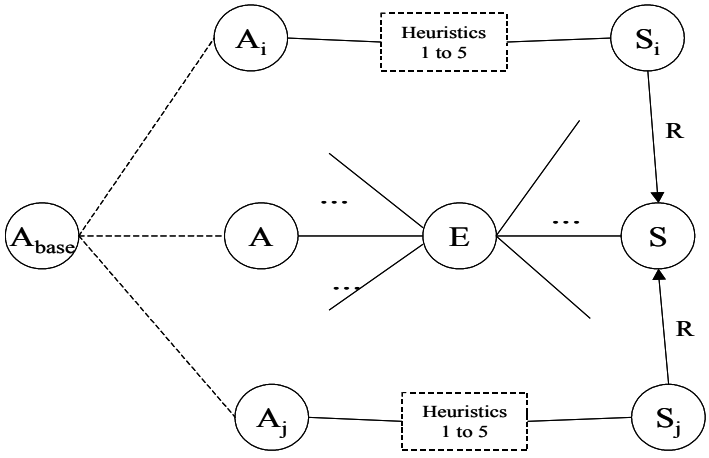

Fig. 5. Graph for heuristic 4

5. Heuristic 5 is the same as heuristic 4 except that there are multiple translations E_1, E_2, … of A and, for each translation E_i, there are possibly multiple associated synsets S_{i1}, S_{i2}, …. In this case the output tuple <A-S> is tagged as 5 and again the heuristic can be sub-classified by the number of input edges or supporting semantic relations (1, 2, 3 ...). See Figure 6.

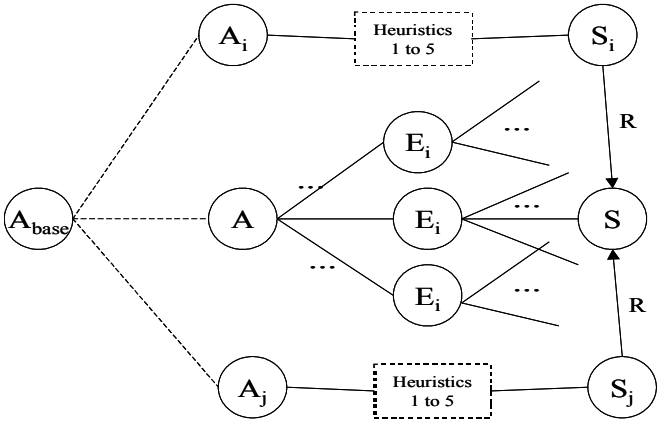

Fig. 6. Graph for heuristic 5

3.2.5 A Detailed Example

Consider once more the case of verb درس (DaRaSa, to learn). From the 9 forms obtained by applying Rule Set 1 to the basic form, the filter accepts the Classes 2, 4 and 7 (as shown in Table 3 on p.7 above). Here we look at the basic form and the Class 2 derivate. We begin by collecting the following tuples using the NMSU dictionary and PWN:

درس:	learn:	['00578275', '00579325', '00584743', '00580363', '00801981', '00890179']:verb
درس:	study:	['00623929', '00587590', '02104471', '00580363', '00587299', '00681070']:verb
دَرَّس:	instruct:	['00801981', '00725200', '00803912']:verb
دَرَّس:	teach:	['00801981', '00264843']:verb
درس:	teach:	['10599680']:noun
درس:	study:	['00608171', '05422945', '06775158', '05374971', '04177786', '05644624', '04065428', '05450040', '09971266', '06616749']:noun
درس:	lesson:	['00836504', '06262123', '06198025', '00686199']:noun
مدروس:	studied:	['01738792', '01782596']:adjective
دارس:	researcher:	['09837494']:noun
دارس:	studying:	['06190701']:noun

دارس:	student:	['09970518', '09869332']:noun
درّس:	study:	['00623929', '00587590', '02104471', '00580363', '00587299', '00681070']:verb
تدریس:	teaching:	['00834401', '05811310', '00831015']:noun
تدریس:	instruction:	['06369463', '00831015', '00834401', '06178338']:noun
تدریس:	faculty:	['05325039', '07787222']:noun
مدرس:	school:	['07776854', '03989548', '05424562', '07777509', '14342474', '07775337', '07512364']:noun
مدرس:	teacher:	['09997151', '05515561']:noun
مدرس:	instructor:	['09997151']:noun

Between the synsets identified above, the following relations hold:

07776854 has as a member 07787222
07787222 is a member of 07776854
00801981 cause 00578275
00686199 is a part of 00831015
00831015 has as a part 00686199
00578275 is a type of 00587299
00587299 has as a type 00578275
00587299 is a type of 00584743
00584743 has as a type 00587299
00834401 is a type of 00836504
00836504 has as a type 00834401

Using these relations, we build an undirected graph where nodes correspond to synsets and edges to semantic relations between synsets. Table 5 shows the 12 candidate associations generated of which 9 are deemed correct by the lexicographers. Note that no candidates have been selected on the basis of the heuristic 1 or heuristic 4. Note also that subclasses of heuristic 5 (rows 9 to 12) are somewhat overvalued because nodes connected by relations with inverses are counted twice.

Table 5: Candidates for Class 1 and 2 derivates of درس (DaRaSa, to learn)

	Buckwalter	POS	Synset Off	Class	Arabic form	Lex Judge
1	drs	verb	580363	2	درس	ok
2	drs	verb	801981	2	درس	ok
3	tdrys	noun	834401	2	تدريس	ok
4	tdrys	noun	831015	2	تدريس	ok
5	mdrs	noun	9997151	2	مدرس	ok
6	drs	noun	836504	3	درس	ok
7	drs	noun	686199	3	درس	ok
8	drs	verb	578275	3	درس	ok
9	drs	verb	587299	5,5	درس	ok
10	drs	verb	584743	5,3	درس	no
11	mdrs	noun	7776854	5,3	مدرس	no
12	tdrys	noun	7787222	5,3	تدريس	no

The first row in Table 5 corresponds to the tuple < 580363 ,درس >. It has been selected on the basis of heuristic 2 because the synset 580363 occurs in:

درس : to learn: […, '00580363', …]

درس : to study: [… , '00580363', …].

The sixth row of Table 5 corresponds to the tuple < 836504 ,درس >. In this case, heuristic 3 can be applied because in

درس : lesson: […, '00836504' ,…]

the synset 00836504 is related to the synset 00834401 by a hyponymy relation:

00836504 has as a type 00834401

which, in turn, has been suggested on the basis of heuristic 2 (see row 3 in Table 5). Finally, consider the tuple <00587299 , درس> in row 9 of Table 5. This is an example of the application of heuristic 5. In

درس : to study: […, '00587299', …]

the synset 00587299 receives support from (among others):

00578275 is a type of 00587299

00584743 has as a type 00587299

where 00578275 and 00584743 have been associated with other derivative forms of درس (DaRaSa, to learn) as shown in rows 8 and 10 respectively of Table 5.

3.2.6 Evaluation

To perform an initial evaluation of this approach, we randomly selected 10 of the 2296 verbs currently in AWN that have a non null coverage and which satisfy all the

requirements above. In addition, for the purpose of illustration, we added the verb درس (DaRaSa, to learn) as a known example. The process for building the candidate set of Arabic form-synset associations described in Section 3.2.4 was applied to each of the 11 basic verb forms resulting in 11 sets of candidate tuples. The size in words and synsets are presented in Table 6.

Table 6: Size of the candidate sets for testing

Arabic form	# of words	# of synsets
عَـامَـلَ	107	190
أَعْقَبَ	71	77
صَقَّلَ	31	21
رَتَّبَ	62	102
أَخَّرَ	19	9
أَخْبَرَ	80	105
رَشَّحَ	40	22
غَامَرَ	56	49
أَشْبَعَ	38	34
أَخْرَجَ	85	140
دَرَّسَ	57	51

Each of the tuples was then scored following the procedure described in Section 3.2.4.4. We did not introduce a threshold and so the whole list of candidates, ordered by reliability score, was evaluated by a lexicographer. The results are presented in Table 7. Here, the first column indicates the scoring heuristic applied, the second the number of instances to which it applied, the third the number of instances judged acceptable by the lexicographer, the fourth the number of instances judged unacceptable, and the fifth the percentage correct.

These results are very encouraging especially when compared with the results of applying the EuroWordNet heuristics reported in Section 3.1. While the sample is clearly insufficient (for instance, there are no instances of the application of heuristic 1 and too few examples of heuristic 3), with few exceptions the expected trend for the reliability scores are as expected (heuristics 2 and 3 perform better than heuristic 4 and the latter better than heuristic 5). It is also worth noting that heuristic 3, the first that relies on semantic relations between synsets in PWN, outperforms heuristic 2. However, we have not attempted to establish statistical significance because of the small size of the test set. Otherwise, an initial manual analysis of the errors shows that several are due to the lack of diacritics in the resources.

Currently we are extending the coverage of the test set. We will then repeat the entire procedure using only dictionaries containing diacritics. We are also planning to refine the scoring procedure by assigning different weights to the different semantic

relations between synsets. In addition, we expect to compare this approach with that based on Bayesian Networks mentioned earlier.

Table 7: Results of the evaluation of proposed Arabic word-PWN synset associations

Heuristic	#	# ok	# no	% correct
1	0	0	0	0
2	42	27	15	64
3	19	13	6	68
4,1	0	0	0	0
4,2	7	4	3	57
4,3	9	5	4	56
4,4	2	1	1	50
4,5	2	1	1	50
4,6	0	0	0	0
4,7	1	0	1	0
5,1	0	0	0	0
5,2	63	32	31	51
5,3	109	41	68	38
5,4	4	4	0	1
5,5	10	6	4	60
5,6	1	1	0	100
5,7	2	0	2	0
5,13	1	0	1	0
Total	272	135	137	50

4 Outlook and Conclusion

We have presented the current state of Arabic WordNet and described some procedures for semi-automatically extending AWN's coverage. On the one hand, the procedure for suggesting translations on the basis of 8 heuristics used for EuroWordNet was presented and discussed. On the other, we described a set of procedures for the semi-automatic extension of AWN using lexical and morphological rules and provided the results of their initial evaluation.

We hope that work will continue on augmenting the AWN database by both manual and automatic means even after the current project ends. We welcome ideas, suggestions, and expressions of interest in contributing or collaborating on both further extension of the lexical database as well as on development of related software. Finally, we are looking forward to a wide range of NLP applications that make use of this valuable resource.

Acknowledgement

This work was supported by the United States Central Intelligence Agency.

References

1. Black, W., Elkateb, S., Rodriguez, H, Alkhalifa, M., Vossen, P., Pease, A. and Fellbaum, C., (2006). Introducing the Arabic WordNet Project, in *Proceedings of the Third International WordNet Conference*, Sojka, Choi, Fellbaum and Vossen eds.
2. Elkateb, S. (2005) *Design and implementation of an English Arabic dictionary/editor*. PhD thesis, The University of Manchester, United Kingdom.
3. Elkateb, S., Black. W., Rodriguez, H, Alkhalifa, M., Vossen, P., Pease, A. and Fellbaum, C., (2006). Building a WordNet for Arabic, in *Proceedings of the Fifth International Conference on Language Resources and Evaluation*, Genoa, Italy.
4. Vossen, P. (ed.) (1998) *EuroWordNet: A Multilingual Database with Lexical Semantic Networks*. Dordrecht: Kluwer Academic Publishers.
5. Rodríguez, H., Climent, S., Vossen, P., Bloksma, L., Peters, W., Roventini, A., Bertagna, F., Alonge, A., (1998). The top-down strategy for building EuroWordNet: Vocabulary coverage, base concepts and top ontology. *Computers and Humanities, Special Issue on EuroWordNet* 32, 117–152.
6. Fellbaum, C. (ed.) (1998) *WordNet: An Electronic Lexical Database*. Cambridge, MA: MIT Press.
7. Niles, I., and Pease, A. (2001) Towards a Standard Upper Ontology. In: *Proceedings of FOIS* 2001, Ogunquit, Maine, pp. 2-9. (See also www.ontologyportal.org)
8. Vossen P. (2004) EuroWordNet: a multilingual database of autonomous and language specific wordnets connected via an Inter-Lingual-Index. *International Journal of Lexicography*, Vol.17 No. 2, OUP, 161-173.
9. Diab, Mona (2005). The Feasibility of Bootstrapping an Arabic WordNet leveraging Parallel Corpora and an English WordNet. *Proceedings of the Arabic Language Technologies and Resources*, NEMLAR, Cairo 2004.
10. Pease, A., (2003) The Sigma Ontology Development Environment, in *Working Notes of the IJCAI-2003 Workshop on Ontology and Distributed Systems*. Volume 71 of CEUR Workshop Proceeding series
11. Buckwalter, Tim (2002). Arabic transliteration. http://www.qamus.org/transliteration.htm.
12. Brihaye, Pierrick (2003). AraMorph: http://www.nongnu.org/aramorph/
13. Farreres, J. (2005) *Creation of wide-coverage domain-independent ontologies*. PhD thesis, Univertitat Politècnica de Catalunya.
14. Benítez, L., Cervell, S., Escudero, G., López, M., Rigau, G., Taulé, M., (1998). Methods and tools for building the Catalan WordNet. In: *Proceedings of LREC Workshop on Language Resources for European Minority Languages*.
15. Agirre, E., Ansa, O., Arregi, X., Arriola, J., de Ilarraza, A. D., Pociello, E., Uria, L., (2002). Methodological issues in the building of the Basque WordNet: Quantitative and qualitative analysis. In: *Proceedings of the first International WordNet Conference*, 21-25 January 2002. Mysore, India.
16. Tufis, D. (ed.) (2004) Special Issue on the Balkanet Project, *Romanian Journal of Information Science and Technology Special Issue* (Volume 7, Num 1-2).

17. Miháltz, M., Prószéky, G., (2003). Results and evaluation of Hungarian nominal WordNet v1.0. In: et al., S. (Ed.), *Proceedings of the Second International WordNet Conference* (GWC 2004). Masaryk University, Brno, pp. 175–180.
18. Barbu, E., Barbu-Mititelu, V. B., (2005). A case study in automatic building of wordnets. In: *Proceedings of OntoLex 20*05 - Ontologies and Lexical Resources.
19. Magnini, B., and Cavaglia, G. (2000) Integrating Subject Field Codes into WordNet. In Gavrilidou M., Crayannis G., Markantonatu S., Piperidis S. and Stainhaouer G. (Eds.) *Proceedings of the Second International Conference on Language Resources and Evaluation*, Athens, Greece, 31 May- 2 June, 2000, pp. 1413-1418.
20. Chen, H., Lin, C., Lin, W., (2002). Building a Chinese-English WordNet for translingual applications. *ACM Transactions on Asian Language Information Processing* 1 (2), 103–122.
21. Farreres J., Rodríguez, H., and Gibert, K. (2002) Semiautomatic creation of taxonomies. SemaNet'02: Building and Using Semantic Networks, in conjunction with COLING 2002, August 31, Taipei, Taiwan See: <http://www.coling2002.sinica.edu.tw/>
22. Wehr, H. (1976) *Arabic English Dictionary*. Cowan.
23. Witten, I.H. and Frank, E. (2005) *Data mining: Practical machine learning tools and techniques* (second edition). San Francisco, CA: Morgan Kaufmann.