On the Evaluation and Improvement of Arabic WordNet Coverage and Usability

Lahsen Abouenour[1], Karim Bouzoubaa[1], Paolo Rosso[2]

[1] *Mohammadia School of Engineers, Mohammed V University-Agdal, Rabat, Morocco*

[2] *Natural Language Engineering Lab., ELiRF, Universitat Politècnica de València, Spain*

abouenour@yahoo.fr, karim.bouzoubaa@emi.ac.ma, prosso@dsic.upv.es

Abstract. Built on the basis of the methods developed for Princeton WordNet (PWN) and EuroWordNet, Arabic WordNet (AWN) has been an interesting project which combines WordNet structure compliance with Arabic particularities. In this paper, some AWN shortcomings related to *coverage* and *usability* are addressed. The use of AWN in Question/Answering (Q/A) helped us to deeply evaluate the resource from an experience-based perspective. Accordingly, an enrichment of AWN was built by semi-automatically extending its content. Indeed, existing approaches and/or resources developed for other languages were adapted and used for AWN. The experiments conducted in Arabic Q/A have shown an improvement of both AWN *coverage* as well as *usability*. Concerning *coverage*, a great amount of Named Entities (NEs) extracted from YAGO were connected with corresponding AWN synsets. Also, a significant number of new verbs and nouns (including Broken Plural forms) were added. In terms of *usability*, thanks to the use of AWN, the performance for the AWN-based Q/A application registered an overall improvement with respect to the following three measures: accuracy (+9.27% improvement), mean reciprocal rank (+3.6 improvement) and number of answered questions (+12.79% improvement).

Keywords. *Arabic WordNet, hyponymy extraction, maximal frequent sequence, WordNet-based application*

1 Introduction

The last decade witnessed experiences in building over 40 wordnets (WNs), aiming for better coverage of main concepts and semantic relations and giving rise to many development methods to overcome several known wordnet challenges. These challenges became more conspicuous when dealing with languages less commonly addressed by Natural Language Processing (NLP) research. The latter case includes, among others, Arabic and Hebrew, the most prominent members of the Semitic family.

Construction of Arabic WordNet (AWN) (El kateb et al. 2006) followed the general trend, leveraging the methods developed for Princeton WordNet (PWN) (Fellbaum 1998) and EuroWordNet (Vossen 1998). The result was a linguistic and semantic resource that complies with the WN structure while considering some specificities of Arabic such as entry vocalization, Broken (irregular) Plurals (BP) and roots. The first release of this resource may well be viewed as a valuable step in terms of the following findings:

- The most common concepts and word-senses in PWN 2.0 have been considered in AWN.
- AWN provides some culture-specific senses. For instance, the word sense أرض الكنانة (The land of Egypt), which is commonly used in Arabic to refer to the country "Egypt", belongs to the synset "جُمْهُورِيَّة" (republic).[1]
- AWN is designed and linked to PWN synsets so that its use in a cross-language context is possible.

[1] In this paper, we use the Buckwalter transliteration (see http://www.qamus.org/transliteration.htm)

- Similarly to other wordnets, AWN is connected to SUMO (Suggested Upper Merged Ontology) (Niles and Pease, 2001; Niles and Pease, 2003; Black et al., 2006). A significant number of AWN synsets was, indeed, linked to their corresponding concepts in SUMO. Statistics show that 6,556 synsets in AWN (65.56% of the synsets) are linked to 659 concepts in SUMO (65.9% out of 1000 concepts). Definitions that are provided by SUMO and its related domain-specific ontologies can be of great interest, complementing the information contained in AWN (SUMO also covers the Arabic culture domain).

Before releasing AWN, the lack of linguistic resources had always been an obstacle to the development of efficient and large scale Arabic NLP systems. Once released, AWN quickly gained attention and became known in the Arabic NLP community as one of the rare freely available lexical and semantic resources.

Nearly five years now since the AWN project was launched, we have found it interesting to evaluate the resource in terms of two aspects: *coverage* and *usability*. Concerning AWN *coverage*, it seems logical to begin by comparing AWN contents with those of a lexicon covering modern standard Arabic and with other wordnets. AWN contains around 18,925 Arabic word-senses[2] belonging to roughly 9,698 synsets,[3] very poor content indeed in comparison to other wordnets. Table 1 presents a comparison among Arabic, Spanish[4] and English[5] WordNets contents, as well as the estimated ratio of the number of word lemmas in each Wordnet to the number of words in large lexical resources corresponding to each language.[6]

Table 1 Comparison of AWN content with an Arabic lexicon and other WNs

Figures	Arabic	Spanish	English
WN Synsets	9,698	57,424	117,659
WN Word-Senses	18,925	106,566	206,941
WN Word Lemmas (WL)	11,634	67,273	155,287
Language Lemmas (LL)	119,693	104,000	230,000
Ratio lemmas (WL/LL)	9.7%	64.7%	67.5%
Ratio Word-lemmas (WN/English WN)	7.5%	43.3%	100.0%
Ratio Synsets (WN/English WN)	8.2%	48.8%	100.0%
Ratio Word-senses (WN/English WN)	9.1%	51.5%	100.0%

Table 1 shows that (i) on the one hand, the released AWN contains only 9.7% of the estimated number of word lemmas in the Arabic lexicon considered (versus 67.5% for the English WN and 64.7% for the Spanish WN), which in turn represent roughly 7.5% of those existing in English WN; and (ii) on the other hand, the number of synsets in AWN represents only 8.2% of the English WN synsets (versus 48.8% for Spanish WN).

The link between word lemmas and synsets is established through word-sense pairs that represent 9.1% of what exists in English WN (51.5% in the case of Spanish WN). Furthermore, AWN synsets are linked by only three kinds of relations (hyponymy, synonymy and equivalence), versus the seven semantic relations used in English WN (which also include antonymy and meronymy, among others).

In previous work (Alotaiby et al. 2009), experiments conducted on nearly 600 million tokens from the Arabic Gigaword corpus (Graff 2007) and the English Gigaword corpus (Graff et al. 2007) showed that the total number of Arabic word types needed in any application is 1.76 times greater than that of English word types required for the same application. On the basis of the foregoing statistics, it is clear that AWN *coverage* is limited compared to the DIINAR.1 lexicon

[2] In WordNet, a word lemma that appears in *n* synsets has *n* word-senses.
[3] AWN statistics are extracted from the AWN browser and database available at: http://www.globalwordnet.org/AWN/AWNBrowser.html
[4] Spanish WN 1.6 statistics are extracted from the MultiWordNet project, see: http://multiwordnet.fbk.eu/online/multiwordnet-report.php
[5] English WordNet 3.0 statistics are extracted from: http://wordnet.princeton.edu/wordnet/man/wnstats.7WN.html
[6] The considered lexical resources are: DIINAR.1 lexicon for Arabic which presents the advantage of containing voweled and lemmatized entries that exist in the language, the Spanish lexicon and the British English Source Lexicon (BESL) for English (both are large and contain morphological information). The three resources are published by ELRA (statistics are extracted from http://catalog.elra.info).

for Arabic and to other WNs. Therefore, one may question the usefulness of the resource and its response to the needs of different applications.

Another point that deserves mention is AWN *usability*. While the efficacy of other WNs (e.g., English and Spanish) in different sophisticated NLP applications has been proven through several research efforts and experimental results (Kim et al. 2006; Wagner 2005), AWN was considered in just a few applications. In fact, AWN was only used and cited as:

- a comparative resource to evaluate a Web-based technique for building a lexicon from hypernymy relations with hierarchical structure for Arabic (Elghamry 2008);
- a resource for Query Expansion (El Amine 2009);
- a resource to be linked to the PanLex 2.5 which is a database that represents assertions about the meanings of expressions (Baldwin et al. 2010); [7]
- a source of information for building an Arabic lexicon by incorporating traditional works on Qur'anic vocabulary (Sharaf 2009);
- a promising resource that (i) allows the exploration of the impact of semantic features on the Arabic Named Entities Recognition (NER) task (Benajiba et al. 2009) and (ii) improves the question analysis module in the Arabic Question/Answering (Q/A) system called QASAL (Brini et al. 2009a; Brini et al. 2009b).

In summary, AWN presents many advantages, including WN structure compliance, mapping to other ontologies and consideration of some Arabic specificities; nevertheless, its patent coverage weaknesses explain its use in just a few projects. Currently, world-wide interest in the development of WNs is increasing. As a matter of fact, the last edition of the Global WordNet conference[8] revealed around 55 projects related to new WN construction, existing WNs enrichment, WNs and lexical resources integration, WN applications and other WN efforts. The AWN project will have to keep up with such dynamism.

The goal of this research is therefore to contribute to the development of a second release of AWN by enhancing its *coverage* and promoting its *usability* in the context of an Arabic Q/A system. The work is three-fold:

- The first phase of this research deals with AWN *usability* in Arabic Q/A;
- The second phase consists in analyzing the inefficiency of using AWN for Arabic Q/A;
- The third phase is an extension of AWN *coverage*.

Jointly, the three phases aim to explore different possibilities for extending AWN coverage so as to increase the usefulness of AWN for Arabic NLP in general, while satisfying the specific need to achieve the best performance possible for Arabic Q/A.

This paper is organized as follows: Section 2 analyzes AWN weaknesses. It also presents a resource-based and a process-based extension of AWN content and It ends by giving a summary of the observed coverage gains. Section 3 highlights how AWN was integrated into a Query Expansion (QE) process used in an Arabic Q/A application; then, it presents the new achievements after the AWN extended version has been used. Finally, in Section 4, the main conclusions of our work are drawn and a list of some future works is previewed.

2 Semi-automatic Extension of AWN *Coverage*

In order to address the main lines to be followed in extending AWN *coverage* for promotion of its *usability*, a detailed analysis of AWN content is required. There is also a need to identify the gap between this content and what is required by NLP applications, such as Arabic Q/A, in terms of resource coverage. The first part of this section presents an analysis of AWN content undertaken on the basis of various statistics. The second part explains how semi-automatic extension can be performed through both resource-based and process-based approaches.

[7] http://utilika.org/info/panlex-db-design.pdf

[8] The conference has been held every two years since 2004. The most recent was the 2012 edition (http://lang.cs.tut.ac.jp/gwc2012/).

2.1 Analyzing AWN Weakness

To make the AWN coverage described in Table 1 more precise, detailed figures about the number of AWN synsets and words are presented in Table 2 with an emphasis on the following three elements:

- Nouns and verbs, as the main Common Linguistic Categories (CLC);
- Named Entities (NEs), as one of the most important types of dynamic information to link with the AWN resource, since AWN is designed for various Arabic NLP applications and domains, including the Web, where NEs are widely used;
- Broken plurals, as a linguistic characteristic mainly specific to Arabic, which are formed by changing the word pattern, not by using regular suffixation. AWN can be used in different NLP applications, particularly, in Information Retrieval, but the Arabic light stemming algorithms that are reported to be effective in this field do not extract the correct stem for BP (Goweder and De Roeck 2001). The use of lexical resources that integrate these BP forms can resolve such problems. It makes sense therefore to devote more attention to the enrichment of AWN in terms of BP forms.

Table 2 Detailed AWN statistics

Figures	CLC		Dynamic information	Arabic-specific characteristic
	Nouns	Verbs	Named Entities	Broken Plurals
No. AWN Synsets	7,162	2,536	1,155	126
No. AWN Word-senses	13,330	5,595	1,426	405
No. AWN Distinct Lemmas	9,059	2,575	1,426	120
No. Baseline Lexicon Lemmas (BLL)	100,236	19,457	11,403	9,565
Percentage AWN Lemmas/BLL	9.0%	13.2%	12.5%	1.3%

In Table 2, we compare the number of lemmas in AWN with DIINAAR.1 as a baseline lexicon (Abbès et al. 2004). This comparison shows that, with respect to the three elements under consideration (CLC, Dynamic Information, etc.), the gap between the two lexical resources is significant. In fact, lemmas in AWN account for only around 9% of nouns and 13.2% of verbs in the baseline lexicon. For dynamic information, this percentage is about 12.5%. The BP forms, peculiar to Arabic, are hardly covered in AWN: it only contains 1.25% of similar forms in the baseline lexicon.

In previous work (Abouenour et al. 2009a), detailed in Section 3, we were interested in the *usability* of AWN for Arabic Q/A systems. AWN helped us to improve the quality of passage ranking. For each user question, the underlying process tries to retrieve passages from the Web most likely to contain the expected answer. Our process is mainly based on a Query Expansion (QE) module which is applied to each question keyword. This module works following two steps: (i) the identification of the AWN synsets that concern the given keyword; and (ii) the extraction of new terms semantically related to the given keyword from AWN. Consequently, the overall performance of the AWN-based approach will be impacted by two factors: (i) non-coverage of question keywords by AWN, so that the first step can not be applied, and (ii) extraction, in the second step, of a limited number of related terms. In order to evaluate AWN in relation to these two factors, we analyzed 2,264 translated questions extracted from CLEF[9] and TREC.[10] The results obtained are given in Table 3. Note that the figures of the last four rows of the table were manually calculated.

Data in Table 3 show that we were able to apply the AWN-based QE process to only 65% of the questions considered in that study—the remaining 35% contained keywords that were not covered by AWN—and that the keywords covered can be expanded by, on average, 4 corresponding synonyms from AWN.

[9] Conference and Labs of the Evaluation Forum: http://www.clef-campaign.org
[10] Text REtrieval Conference: http://trec.nist.gov/data/qa.html

Table 3 Analysis of the AWN coverage for the CLEF and TREC questions

Indicators	CLEF	TREC	Overall	%
No. Questions	764	1,500	2,264	-
No. Questions covered by AWN	612	858	1,470	64.93%
Avg. AWN word lemmas per question	3.65	4.26	4	-
No. Questions Not Covered (QNC) by AWN	152	642	794	35.07%
QNC with NE keywords	127	420	547	68.89%
QNC with Verb keywords	44	262	306	38.54%
QNC with Noun keywords	81	508	589	74.18%
QNC with Broken Plural keywords	0	18	18	2.27%

A more in-depth analysis of the results in Table 3 reveals that over 74% of the questions not covered by AWN contain noun word lemmas, around 69% include NEs and roughly 39% are composed of at least one verb. We can also notice that BP forms (the irregular form of plural) are present in over 2% of these questions (only 120 such forms exist in AWN: this represents around 1.71% of the well-known existing BP lists). For example, the TREC question " متى وقعت حرائق الرايخستاغ ؟" (When did the Reichstag fires happen?) is formulated with three keywords: the verb "وقع" (happen), the BP "حرائق" (fires) and the NE "الرايخستاغ" (Reichstag). Since none of these keywords exists in AWN, the question can not be extended using the QE process.

The figure from our Q/A study displays the AWN weaknesses previously pointed out and highlights the need to expand its coverage. To extend AWN content, particular interest was attached to semi-automatic methods among the most commonly used by researchers when enriching wordnets. These methods help to avoid the limitations of: (i) the manual approach, which consumes time and effort and tends to result in low coverage resources; and (ii) the automatic approach, which raises the coverage to the detriment of accuracy and confidence. In the following subsections, we propose two types of AWN extension: (i) Resource-based extension of NEs and verbs using existing English resources, and (ii) Process-based extension of nouns using a hyponymy pattern recognition process. The fact that the second extension is process-based explains why the corresponding subsection is more detailed.

2.2 Resource-based AWN extension

Diab (2004) already proposed a resource-based AWN extension by means of Arabic English parallel corpora and English WordNet. In this subsection, we also extend AWN on the basis of existing English resources. Rather than using parallel corpora in recovering the Arabic side, we have explored using the Google Translation tool which can provide good results when processing unique entries (NEs or verbs).

2.2.1 Named Entities Extension using the YAGO Ontology

Various research efforts have aimed at extending wordnets with NEs. Indeed, adding new NEs synsets to WN is of paramount importance in the field of NLP because it allows using this unique resource for NE recognition and other tasks. Toral et al. (2008) automatically extended PWN 2.1 with NEs using Wikipedia. NEs in Wikipedia are identified and integrated in a resource called Named Entity WordNet, after a mapping performed between the is-a hierarchy in PWN and the Wikipedia categories. Al Khalifa and Rodriguez (2009) also demonstrated that it is possible to enrich NEs in AWN by using the Arabic Wikipedia: in that work, experiments showed that 93.3% of automatically recovered NE synsets were correct. However, due to the small size of the Arabic Wikipedia, only 3,854 Arabic NEs could be added.

One way to tackle monolingual resource scarcity problems is to use available resources in one language to extend existing WordNet in another, as was done by Benoît and Darja (2008) for French WN.

In a previous work (Abouenour et al. 2010b),[11] we proposed a technique that allows enriching the NE content in AWN on the basis of the large English NE ontology called YAGO[12] (Suchanek

[11] This work was conducted under the framework of the bilateral Spain-Morocco research project AECID-PCI C/026728/09 (PI Horacio Rodriguez, Technical University of Catalonia).

[12] Yet Another Great Ontology: available at http://www.mpi-inf.mpg.de/YAGO-naga/YAGO/downloads.html

et al. 2007). In fact, the high coverage of NEs in YAGO (around 3 million), the claimed 95% accuracy, the mapping with WordNet, the connection with SUMO and further advantages have led us to investigate the degree to which it would be useful to translate the content of YAGO into Arabic and integrate it into AWN. The proposed technique is composed of three steps:
(i) The translation of YAGO entities into Arabic instances by means of Google Translation API (GTA).[13] Based on the manual checking of 1,000 translated NEs, we have observed that this automatic translation has attained an accuracy of 98.2% when applied to a one or two-word NE.
(ii) The extraction of candidate AWN synsets to be associated with the created instances. It was possible to add the translated YAGO entities to AWN through two kinds of mappings:

- Firstly, the WordNet synsets corresponding to a given YAGO entity are extracted using the facts involving the YAGO "TYPE" relation (in YAGO, there are 16 million facts for this relation); the AWN synsets corresponding to the identified WordNet synsets are then connected with the given entity. For example, the YAGO entity "Abraham_Lincoln" appears in three facts for the YAGO "TYPE" relation; from these facts, the three English WN synsets "president", "lawyer" and "person" are extracted. Hence, the YAGO entity "ابراهام لينكولن" (i.e., Abraham Lincoln) can be added as an instance corresponding respectively to AWN synsets identified by "**رَئِيس**" (president), "**مُحَام، مُحَامِي، وَكِيل**" (lawyer, attorney) and "**شَخْص، إنْسَان**" (person, human);
- The second kind of mapping consists in supposing that the arguments of some YAGO relations can be systematically added to AWN as instances of specific synsets. For example, the second argument of the YAGO relation "bornIn" is likely to be an instance of the AWN synset "مدينة" (city : identified by madiynap_n1AR in AWN). Following this idea, we have specified for a set of 19 YAGO relations (out of 99) whether the first or the second argument of the relation should be used and which AWN synset to link should be linked to it. Using this mapping, 331,851 candidate NEs have been extracted and passed on to the validation process.

(iii) The automatic validation of NE links to corresponding AWN synsets. This step aims at eliminating incorrect mappings as well as wrongly translated entities. For instance, in YAGO, the entity "Association_for_Computing_Machinery" is present in the second argument of the relation "isLeaderOf". Therefore, with respect to the evident mapping (the first kind described in (ii) above), this entity is a candidate for being an instance of the synset بلد (country : balad_n1AR). Using the Yahoo API, we extract the Web snippets that match the exact expression " بلد جمعية الآلات الحاسبة" (Association for Computing Machinery country). The given entity is then added in the AWN extension only if the number of extracted snippets exceeds a specific threshold (set heuristically to 100).

After applying this technique on the three million YAGO entities, we found out that it was possible to keep 433,339 instances (145,135 NEs thanks to the first mapping and 288,204 NEs from the second mapping) that were connected with 2,366 corresponding AWN synsets. This number represents around 38,000 times the number of existing NE instances in AWN. Table 4 presents statistics of NE classes that were augmented in AWN.

Table 4 Statistics of NE classes augmented in AWN

Cat. ID	NE categories	Number	%
1	PERSON	163,534	37.7%
2	LOCATION	73,342	16.9%
3	EVENT	14,258	3.3%
4	PRODUCT	14,148	3.3%
5	NATURAL OBJECT	8,512	2,0%
6	ORGANIZATION	8,371	1.9%
7	FACILITY	4,312	1,0%
8	UNIT	3,513	0.8%
	Sub Total	289,990	66.9%
9	OTHER	143,348	33.1%
	Total	433,339	100%

[13] http://code.google.com/p/google-api-translate-java/

As shown in Table 4, 66.9% of the NEs that were linked to AWN synsets are classified under 8 categories. The most frequent are PERSON (37.7%) and LOCATION (16.9%). The remaining NEs (33.1%) are grouped under the OTHER category.

Most of the added PERSON entities are foreign names; however, this will not impact the experimental process using TREC and CLEF questions containing the same nature of names. Also, we did not investigate using an Arabic NER system as alternative to the resource-based approach so as to avoid any eventual inaccuracy of such a system.

The feasibility of enriching AWN coverage by NEs coming from YAGO was investigated. Nevertheless, we understand that building an Arabic YAGO linked to the English one could presumably be the most suitable option for dynamic information such as NEs. The interesting amount of NEs that we have linked to AWN synsets will at least help in considering their mapping to already existing PWN NEs.

2.2.2 Verb Content Extension using VerbNet and Unified Verb Index

Rodriguez et al. (2008a) have investigated two possible approaches for extending AWN. In both cases, purpose was just to show the potential usefulness of such approaches for semi-automatic extension of the resource. In both works, it was reported that the results were very encouraging, especially when compared with the results of applying the eight EuroWordNet heuristics (Vossen 1998). However, further experiments are needed in order to add number of words to AWN synsets. The first approach deals with lexical and morphological rules, while the second considers Bayesian Network as an inferencing mechanism for scoring the set of candidate associations (Rodriguez et al. 2008b). The Bayesian Network doubles the number of candidates of the previous heuristics approach (554 vs. 272).

In our own work, in order to enrich the verb content in AWN, we have followed a two-step approach inspired by what was proposed by Rodriguez et al. (2008a). The first step consists in proposing new verbs to add to AWN; the second step aims at attaching these newly proposed verbs to corresponding AWN synsets.

Considering the first step, while Rodriguez and his colleagues made use of a very limited but highly productive set of lexical rules in order to produce regular verbal derivative forms, we got these forms by translating the current content of VerbNet (Kipper-Schuler 2006) into the Arabic language. Our reasons were two-fold:

(i) To avoid the validation step where we need to filter the noise caused by overgeneration of derivative verb forms (unused forms can be generated);

(ii) To allow advanced AWN-based NLP applications to use the syntactic and semantic information about verb classes in VerbNet and their mappings to other resources such as FrameNet (Baker et al. 2003) and PropBank (Palmer et al. 2005).

The translation concerned the 4,826 VerbNet verbs distributed into 313 classes and subclasses. After the process of translating every single verb using the Google Translation Web page (note that, unlike GTA, this translation Web page can provide more than one possible translation for a unique verb entry), a manual validation was performed to check the correctness of the translation, and to choose the verb lemmas to be added to AWN. Thanks to this semi-automatic process, we were able to have 6,654 verbs for consideration in the next step. The same process was applied on verbs covered by the Unified Verb Index (UVI).

In the second step, the attachment of Arabic verbs with AWN synsets was done by setting a graph which connects each Arabic verb with the corresponding English verbs that are present in PWN. Figure 1 illustrates this step: A stands for the Arabic verb, Ej for the English verb number j, Si for PWN synset number i and Sai for AWN synset number i.

As Figure 1 shows, each English verb can be connected to different PWN synsets. Then they are connected with their equivalent synsets in AWN. After building the graph connecting each Arabic verb with the corresponding PWN synsets through English verbs, the relevant connections were selected by applying 3 of the 5 graph heuristics adopted in (Rodriguez et al. 2008a). We set the limit at the third heuristic because the percentage of noise attachment increases starting from the fourth heuristic and even more after applying the fifth one.

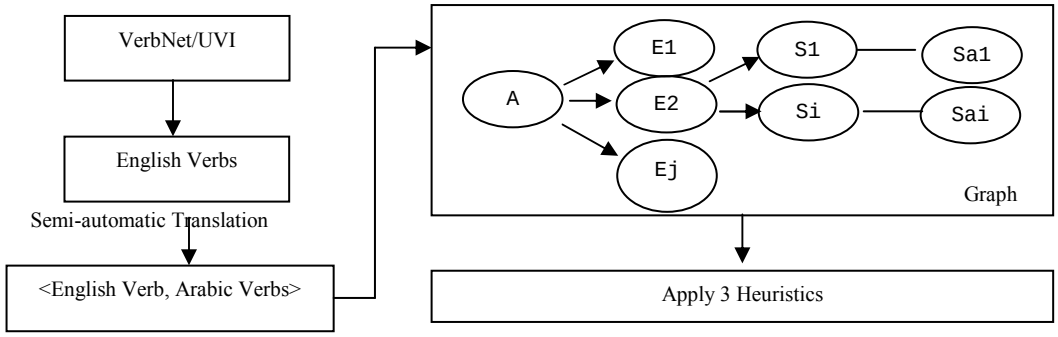

Fig. 1 Enrichment of verbs in AWN and their attachment to synsets

Let us recall the definition of each heuristic as described in that work:

- Heuristic 1: If a unique path AES 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;
- Heuristic 2: If multiple paths AE1S and AE2S exist (i.e., A is translated as E1 or E2 and both E1 and E2 are associated with S among other possible associations) then the output tuple <A,S> is tagged as 2;
- Heuristic 3: If S in AES has a semantic relation to one or more synsets, S1, S2 … 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;
- Heuristic 4: If S in AES has some semantic relation with S1, S2 … where S1, S2 … 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;
- Heuristic 5: Heuristic 5 is the same as Heuristic 4 except that there are multiple translations E1, E2, … of A and, for each translation Ei there are possibly multiple associated synsets Si1, Si2, …. In this case the output tuple <A, S> is tagged as 5.

Note that tags 1, 2 and 3 help in identifying the <A, S> tuple generated by the first, second and third heuristic respectively. Table 5 presents the results obtained using the described verb extension process.

Table 5 Results of the AWN verb extension process

	VerbNet		UVI		Total
	Number	Percentage	Number	Percentage	
Considered Arabic verbs	6,654	-	3,431	-	10,085
Connected Arabic verbs	5,329	80.09%	1,115	31,13%	6,444
Verbs existing in AWN	2,760	41.48%	542	15,80%	3,302
Newly Added Verbs (NAV)	2,569	38.61%	573	16,70%	3,142
- NAV with Heuristic 1	184	2,77%	129	3,76%	313
- NAV with Heuristic 2	158	2,37%	43	1,25%	201
- NAV with Heuristic 3	2,227	33,47%	401	11,69%	2,628
Connected AWN synsets	1,361	-	1,906	-	3,267

As we can see from Table 5, our process succeeded in connecting 5,329 of the Arabic verbs translated from VerbNet with the corresponding AWN synsets (1,361 distinct synsets). Even though around 41.5% of these verbs (2,760 verbs) already existed in the current release of AWN, the process added new synset attachments for them. The remaining 2,569 verbs were not in AWN and could be added. Heuristic 1 allowed the generation of a few but accurate verbs and attachments (2.77%), whereas Heuristic 3 succeeded in coming up with a higher number of less relevant verbs (33.47%). With respect to the verbs generated from UVI, the overall newly connected verbs were 6,444, 3,142 of which were new additions.

2.3 Process-based AWN extension

Relying on resource-based extension is not the only line of investigation for enriching wordnets. Process-based semi-automatic techniques have also been adopted by researchers in order to refine the hyponymy relation in wordnets, as well as to add new noun and verb synsets (Hearst 1992; Costa and Seco 2008; Tjong Kim Sang and Hofmann 2007). Hyponymy discovery is another useful direction for wordnet enrichment that allows the automatic extraction of hyponym/hypernym pairs from text resources such as the Web. For instance, A and B form a hyponym/hypernym pair if the meaning of B covers the meaning of A and is broader (Tjong Kim Sang and Hofmann 2007). There have been many attempts aimed at automatic acquisition of such hyponymy pairs. Hearst (1992) was among the first researchers to have proposed and investigated a pattern-based approach in order to resolve this problem. This approach consists mainly in using a set of lexical and syntactic patterns to generate a list of concepts linked using the considered semantic relation. For instance, in English, the pattern "X including Y1 (, Y2, ..., and |or Yn)" helps to identify the nouns Y1, ..., Yn as candidate hyponyms of the noun X. For example, "cinema" and "drawing" can be extracted as hyponyms of "arts" from the text "The institute focuses on different arts including cinema and drawing". It was reported that adopting these kinds of pattern-based approaches allows the harvesting of semantic relations in general and hyponymy particularly in languages such as English (Pantel et al. 2006; Snow et al. 2005), Spanish (Ortega-Mendoza et al. 2007) and Dutch (Tjong Kim Sang and Hofmann 2007).

As for Arabic, there have been few such attempts in comparison to other languages like English. The work of Elghamry (2008), which proposed an unsupervised method to create a corpus-based hypernym/hyponym lexicon with partial hierarchical structure, is one of these few attempts. In that work, the acquisition process was bootstrapped relying on the lexico-syntactic pattern " بعض X مثل Y1…Yn" (some X such as Y1,…Yn). The effectiveness of the suggested method was demonstrated through a comparison between the extracted entries with those of AWN, but a single lexico-syntactic pattern (" بعض X مثل Y1…Yn") was used. This limitation had two causes: (i) it was reported that Arabic patterns which are equivalent to those proposed in (Hearst 1992) do not give significant results and (ii) there was no Arabic parser available to facilitate the detection of noun phrases in the context of the other patterns. With the availability of Open Source Arabic syntactic parsers like the Stanford Arabic Parser,[14] the latter reason is no longer valid: such syntactic parsers can reduce the noise generated by a long list of Arabic lexico-syntactic patterns.

In line with the above-mentioned research efforts for Arabic and other languages, our aim is to augment the coverage of AWN noun synsets (currently there are 7,162 noun synsets versus 82,115 in English WN) while simultaneously enriching the hyponymy (is-a) relation between these synsets. The two-step method proposed by Ortega-Mendoza et al. (2007) and García-Blasco et al. (2010) was adapted to achieve the target enrichment. Figure 2 illustrates the general architecture of our approach.

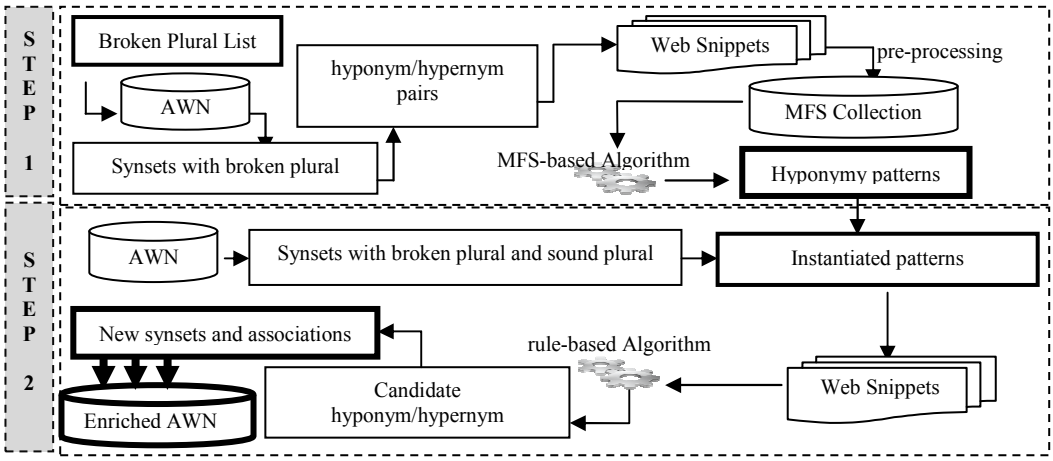

Fig. 2 General architecture for Arabic Hyponym/Hypernym pairs detection

[14] http://nlp.stanford.edu/software/lex-parser.shtml

Figure 2 depicts the two-step method. It can be summarized as follows:

- Step 1: Identify hyponymy patterns over snippets retrieved from the Web. These snippets match a set of queries formed by hypernym/hyponym pairs;
- Step 2: Instantiate the identified patterns. The instantiation is performed by searching for hypernym/hyponym pairs that match the given pattern.

The following subsections present how these steps have been implemented for the Arabic language as well as the results obtained.

2.3.1 Identifying lexico-syntactic patterns

According to Ortega-Mendoza et al. (2007), we need a seed list of hypernym/hyponym pairs to be used as queries. In our case, we have built this list from the synsets existing in AWN. For instance, the synset (fan~ / art) فنّ is described by the following synonyms: (<inotaAj_fan~iy : artistic production) إنْتاج فنّي, (AibodaAE_fan~iy : artistic innovation) إبْداع فنّي and (fan~ / art) فنّ. Figure 3 shows the context of this synset in the AWN hierarchy using the hyponymy relation.

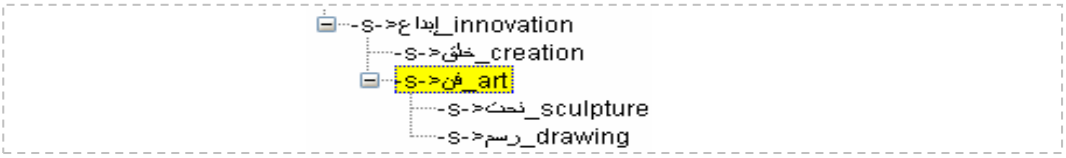

Fig. 3 Context of the synset fan~ in the hierarchy of AWN

As Figure 3 shows, only two hyponyms of the synset (fan~ / art) فنّ are present in the current version of AWN, namely "sculpture" and "drawing". In English WordNet 3.0, 13 hyponyms (gastronomy, perfumery, origami, etc.) exist under the equivalent synset (art).

To know about how this synset appears together with its hyponyms in a text, we have queried the Web with a set of hand-coded hyponymy patterns instantiated using the given synset and its hyponyms. Table 6 describes the used queries and sample snippets obtained as results.

Table 6 Sample snippets obtained using instantiated patterns as queries

Instantiated pattern (in Arabic)	Instantiated pattern (in English)	Sample of obtained snippets	Sample of obtained snippets (in English)
نحت و غير ذلك من **فن**	**sculpture** and other **arts**	فن النحت من أقدم الفنون وأكثرها انتشارًا وتنوعًا في العالم...	Sculpture is one of the oldest arts, the most widespread and diverse in the world...
فن الأخرى خاصة **نحت**	other **art** in particular **sculpture**	الفنون عامة وفن النحت خاصة يعتبر من أهم المجالات التي تعكس بصدق بالغ تفاعلات...	Generally, the arts and in particular sculpture, are one of the most important areas that truly reflect deep interactions...
فن الأخرى على غرار **نحت**	other **arts** such as **sculpture**	قواعد الفن على غرار الفن الاغريقى أو الرومانى...	The rules of art such as Greek or Roman arts...
من أهم هذه **فن** هناك **رسم**	**drawing** is one of the most important **arts**	هناك تقنيات مختلفة للفنون التشكيلية والرسم التي تجعل الاختلافات في الرسم سواء ...	There are different techniques of Fine Arts and painting that make the differences ...

As we can see from Table 6, the hypernym is usually used in its plural form which can be generated by adding specific suffixes (for instance –arts- فنون is the sound plural of فن –art-). This is similar to other languages such as English. According to some research on large Arabic corpora (Goweder and De Roeck 2001; Boudelaa and Gaskell 2002), BP forms constitute around 10% of texts, and BP forms account for 41% of the different plural forms used in texts. Therefore, we used BP forms to automatically extract patterns and built a list of seed hypernym/hyponym pairs starting from the AWN synsets which have a BP form.

Since the current version of AWN contains only a few BP forms, we decided to begin enriching AWN by connecting its synsets and words with such new forms. To perform this task

we relied on 3,000 BP forms extracted from Emad Mohamed's list[15] and automatically connected these forms to the corresponding AWN words using the singular entry existing in that list. The content of the list as well as the connections so-created were manually validated. In all, we connected 1,934 synsets with the corresponding BP form (nearly 24.3% of the AWN noun synsets), using 1,696 hypernym/hyponym pairs to identify lexical patterns (the other synsets do not appear in relevant number of snippets). A description of the procedure used is oulined below.

For each seed pair, we extracted from the Web the first 20 distinct snippets corresponding to the results returned by the Yahoo! API when using the following request forms: "HYPONYM+HYPERNYM" and "HYPERNYM+HYPONYM". The next challenge was to retrieve the relevant lexical patterns from the previously mentioned collection of snippets. Currently, different techniques are suitable for such a task. One of these techniques is based on the retrieval of the Maximal Frequent Sequences (MFS) of words. In fact, many research works (Denicia-Carrel et al. 2006; Ortega-Mendoza et al. 2007; García-Blasco et al. 2010; García-Hernández et al. 2010) highlighted the usefulness of this technique for pattern discovery over text.

Following Ahonen-Myka (2002), a sequence is defined as a set of ordered elements (for instance, words). The frequency of a sequence of words p is determined by the number of sentences that contain p. A sequence is maximal if it is not a subsequence of any other. That is, if it does not appear in any other sequence in the same order. MFS are all the sequences that appear in β sentences (where β is the defined frequency threshold) and are not subsequences of any other MFS. To make these maximal frequent sequences more flexible, García-Hernández (2007) has introduced the concept of gap which is defined as the maximum distance that is allowed between two words in a MFS. Following this, if we set the gap to 0, the words in the MFS will be adjacent words in the original text. For example, $<w_{i0}, \ldots, w_{in}>$, with $i_j \in 1 \ldots k$, is a maximal frequent sequence of k words, $i_j = i_{j-1}+1,\ j > 1$, when $gap = 0$, and $i_j \leq i_{j-1+\eta} + 1$, when $gap = \eta$.

In our work, we adopted MFS for two main reasons: (i) it has achieved higher performance for languages such as English and Spanish (Denicia-Carrel et al. 2006; Ortega-Mendoza et al. 2007; García-Blasco et al. 2010; García-Hernández et al. 2010), and (ii) it is language-independent, which allows us to leverage for Arabic tools that have been developed for the aforementioned languages.

Specifically, we used the MFS-algorithm proposed by García-Blasco et al. (2010). It allows the processing of a document collection (that must be just plain text, divided into lines) and searches for the MFS on the basis of three parameters introduced before running it:

- Minimal Frequency (MF): It is the minimum number of times the sequence must appear. If a sequence appears twice in the same sentence, it will only count as 1 for the frequency;
- Minimal Length (ML): It is the minimum number of words that must compose the sequence;
- Maximal Gap (MG): It is the maximum distance allowed between two consecutive words in the maximal frequent sequence. The greater this value is, the more flexible the extracted patterns will be.

Extracting a high number of hyponymy patterns depends on the coverage of the document collection used. In this work, we built a collection from 102,900 snippets corresponding to 1,696 Web queries (a query is formed from AWN hyponym/hypernym pairs). In order to guarantee the correctness of the extracted patterns, we manually evaluated the patterns that resulted from applying the MFS-algorithm on a small subset of the collection (5,145 snippets, which represent 5% of the collection). We used different parameter values while considering the following constraints: (i) since a MF>20 only generates 2 candidate patterns and a MF<5 generates an excessive number of patterns, we considered a range between 5 and 20 for this parameter, (ii) according to the lengths observed in a manually built list of hyponymy patterns, a range between 3 and 7 was set for MG. Table 7 shows the results of the MFS-algorithm on the small subset of the collection.

As we can see from Table 7, when the parameters are MF=20, ML=2 and MG=7, the algorithm (which is applied on the small subset of the collection) is able to generate 27 candidate patterns of

[15] http://jones.ling.indiana.edu/~emadnawfal/arabicPlural.txt

which 5 patterns (18.52%) are manually qualified as correct hyponymy patterns. This percentage is the highest among the different runs corresponding to the different MFS parameters values.

Table 7 Results of MFS parameter setting in the context of the Arabic language

	Run 1	Run 2	Run 3	Run 4	Run 5	Run 6
Minimal Frequency (MF)	20	20	20	15	10	5
Maximal GAP (MG)	3	5	7	7	7	7
Minimal Length (ML)	2	2	2	2	2	2
#Patterns	19	26	27	46	113	1,019
#Hyponymy Patterns	2	3	5	7	17	135
%Hyponymy Patterns	10.53%	11.54%	18.52%	15.22%	15.04%	13.25%

Now to apply the MFS-algorithm on the whole collection, it makes sense to maintain the same ML and MG parameters values, as they are collection-coverage independent. However, the MF has to be changed to 400. Indeed, unlike ML and MG, the MF depends on the collection coverage and in our case MF is calculated accordingly (MF=102,900*20/5,145). With these parameter values, we succeeded in extracting 23 relevant hyponym patterns from the whole snippet collection. These patterns, after manual validation, were used in the pattern instantiation step (step 2).

2.3.2 Instantiating Patterns

The main objective of the pattern instantiation step is to retrieve candidate hyponym/hypernym pairs with which to enrich the current AWN hierarchy. Generally, a pattern has one of the two following forms: "<Phrase> HYPONYM < Phrase > HYPERNYM" or "HYPERNYM < Phrase > HYPONYM < Phrase >". Instantiating these patterns means that we replace the HYPERNYM part by synset names from AWN and the other parts by a wild character (such as *). For instance, the pattern " العديد من HYPR مثل HYPO" (many HYPR such as HYPO) is instantiated with the synset الأسلحة (Al>slHp : weapons) which is the BP of سلاح (silAH : weapon). The query resulting from this instantiation is: " العديد من الأسلحة مثل *". This query is passed on to the search engine in order to retrieve the most relevant and matching snippets. Table 8 lists samples of the extracted snippets.

Table 8 Sample snippets obtained using the pattern " العديد من HYPR مثل HYPO"

Snippets (in Arabic)	Translation (in English)
وله **العديد من** الأسلحة **مثل**: العصا, السيف... أحد الاسباب التي حدت من انتشار الفن هو التحفظ من المعلمين واختيار التلاميذ بحذر حتى لا تنتقل الاسرار للمنافسين. ...	...have **many** weapons **such as** stick, sword ...
أي معلومات غير موثقة يمكن التشكيك بها وإزالتها. وسم هذا القالب منذ: نوفمبر_2010 ... 1957 حيث تم من خلال تصميمة تطوير وإنتاج **العديد من** الأسلحة **مثل** إم 240 ...	... developing and producing **many** weapons **such as *M240*** ...
تستخدم بعض الأسلحة الكيماوية الغير قاتلة، مثل *الغاز المسيل للدموع ورذاذ* ... فإن **العديد من** الحروب هي جزئيا أو كليا مستندة إلى أسباب اقتصادية، **مثل** الأزمة ...	... several chemical weapons such as *tear gaz* ... **many** wars are completely or partially triggered by economic causes, **such as** *crisis* …
لعام 1939-1945م تجد في اللعبة **العديد من** الاسلحة **مثل** *الدبابات والصواريخ ومدفع الهاون والكثير من* الاسلحة لعبة مثيرة واكشن جداً ...	... you'll find in this game **many** weapons **such as** big *tanks, rockets and mortars* and lot of other weapons ...
هناك **العديد من** الأسلحة **مثل**: *السيوف,الخناجر,الفأس,القوس والسهم,المسدس,تو بليد,إلخ*..... آعذروني لعدم وجود صور للاسلحه آستعراض بسيط للاسطورهـ ...	... There are **many** weapons **such as** *swords, daggers, ax, bow and arrow, pistol* ...
لم تكن تستطيع الوصول إليه من قبل, القتال سيكون باستخدام **العديد من** الأسلحة **مثل** *السوط والسيف والخناجر والفؤوس والسحر* والكثير من الأسلحة الأخرى المتنوعة ...	...using **many** weapons **such as** the *whip and the sword, daggers, axes and magic* ...

In Table 8, the words of the pattern are in bold, the synset used for its instantiation is underlined while the candidate hyponyms are both underlined and in italic. As we can see, in the above example, the left side of the pattern contains the targeted hyponyms. Therefore, a rule-based

algorithm was applied in order to analyze the left side and extract from it nouns that could be added as hyponyms of the synset الأسلحة.

The list of the 23 hyponymy patterns identified in the previous step was instantiated using both 700 AWN synsets (hypernyms) that have BP forms and then using 700 other AWN synsets with their Sound Plural (SP) form. Let us recall that only BP forms have been used as seed pairs of the hyponymy relation while we used both forms in the instantiation phase. This should allow us to determine whether the patterns discovered using a plural form (in our case BP) can be useful in identifying hyponyms for the other form (e.g. SP). Table 9 presents the results obtained.

Table 9 Experimental results of the AWN noun hyponymy extension

Measures	Using BP	Using SP	Overall/Total (distinct)
#AWN hypernym synsets	700	700	1,400
#Successful patterns	17 (73.91%)	9 (39.13%)	17 (73.91%)
#Candidate hyponyms	1,426	828	2,254
Avg. candidate hyponyms per AWN synset	2.04	1.22	1.61
#Correct hyponyms	458 (32.12%)	415 (50.12%)	832 (36.91%)
#AWN hypernym synset with correct hyponyms	94 (13.43%)	191 (27.29%)	284 (40.57%)
#New correct hyponyms (not existing in AWN)	265 (57.86%)	205 (49.40%)	459 (55.17%)
#New AWN associations (hypernym/hyponyms)	193	196	359

As depicted in Table 9, instantiating the 23 patterns with BP forms opens up the possibility of getting an average of around 2 candidate hyponyms per AWN hypernym synset (versus 1.22 using the sound plural form). Note that candidate hyponyms are extracted using a set of automatic rules. These candidate hyponyms are then manually validated in order to identify correct hyponyms (2 persons validated around 2,300 hyponyms within approximately two days). With regard to BP forms, around 74% of the patterns considered succeeded in generating correct hyponyms. The list of these patterns also includes all the patterns that succeeded with SP forms (9 patterns). The difference in pattern accuracy can be explained by the following fact: when using the SP form in the query, snippets often contain the singular instead of the plural stem. Therefore, such snippets will not be relevant and hardly match the pattern considered. For the BP, the program happens not have this confusion.

The results listed in Table 9 also show that 832 correct hyponyms were identified (roughly 37% of the candidate hyponyms). About 60% of these could be added to AWN as new synsets. Even though the remaining hyponyms already existed in AWN, new hypernym/hyponym associations in which they participate could still be added.

According to Table 9, our process succeeded in generating hyponyms for approximately 41% of the 1,400 hypernym synsets considered. The number of hyponyms per hypernym ranges from 1 to 29. Figure 4 illustrates the distribution of the number of hyponyms per hypernym.

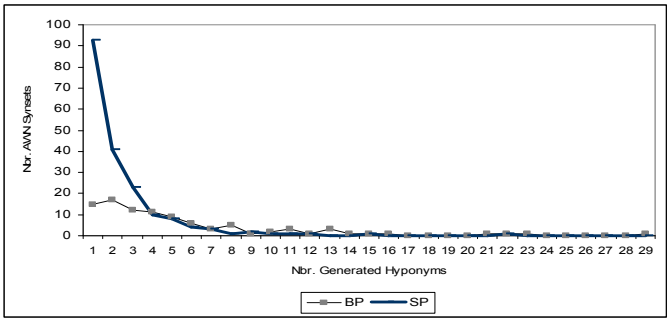

Fig. 4 Distribution of the number of hyponyms per hypernym

Figure 4 contains two curves, corresponding to BP and SP hyponym generation respectively. The first curve reveals that with the BP form, for instance, only one hyponym is extracted for 15 AWN hypernym synsets. While Table 9 shows that SP forms help in generating correct hyponyms for a higher number of AWN synsets (191 vs 94 with BP forms), Figure 4 depicts an unbalanced distribution of these hyponyms over these synsets. In fact, for around 54% of the BP forms the process succeeded in generating at least 4 correct hyponyms, whereas this percentage did not exceed 17.5% for SP forms. To sum up, using both forms as hypernyms guarantees that more

AWN synsets will acquire hyponyms, but not with the same accuracy. Table 10 lists the patterns that generate a high average of hyponyms per synset.

Table 10 Top relevant hyponymy patterns

Pattern	English translation	Avg. hyponyms per synset
HYPO مثل HYPR من العديد	Many HYPR such as HYPO	1.32
HYPO ك HYPR من العديد	Many HYPR for instance HYPO	1.30
HYPO مثل HYPR بعض	Some HYPR such as HYPO	1.13
HYPO مثل الأخرى HYPR	Other HYPR such as HYPO	1.10
HYPO ك الأخرى HYPR	Other HYPR for instance HYPO	0.89
HYPR وغير ذلك من HYPO	HYPO and other HYPR	0,88

As shown in Table 10, the best hyponym patterns contain the hypernym part in the middle or at the beginning. The experimental results show that we have reached our aim, i.e. to enrich the noun content and hierarchy of the AWN. Indeed, thanks to the use of a set of automatically discovered patterns (via an MFS-based algorithm), it was possible to add 459 new synsets (which account for 7.53% of the number of existing noun synsets) and 359 new associations between synsets using the hyponymy relation (around 2% of the existing associations).

The proposed technique is promising since it allows suggesting candidate hyponyms that can be validated and integrated under AWN synsets. In principle, this way is faster than adding these hyponyms from scratch, especially if we consider the following further possibilities:

- Extracting new patterns by setting other values for MFS parameters. These patterns can help in generating new hyponyms;
- Using a recursive process in which generated hyponyms play the role of hypernyms.

Since the technique is relation-independent, it can also be used for enriching AWN by adding new relations between synsets such as the meronymy (part of) relation.

2.4 *Coverage* of the Enriched AWN Resource

As described above, it is possible to semi-automatically extend the content of NEs, verbs and nouns in AWN. For each case, we made use of and adapted existing approaches and/or resources developed for other languages. Thanks to this extension process, we obtained the results summarized in Table 11 and Table 12.

Table 11 Nouns, verbs and NEs *Coverage* improvement

Figures	Common Linguistic Categories			Dynamic Information		
	Nouns and Verbs			Named Entities		
	Original	Extended	Added	Original	Extended	Added
No. AWN Synsets	9,698	10,198	5.2%	1,155	433,339	37,418.5%
No. AWN Word-senses	18,925	37,463	98.0%	1,426	433,339	30,288.4%
No. AWN Distinct Lemmas	11,634	15,005	29.0%	1,426	433,339	30,288.4%
No. Baseline Lexicon Lemmas (BLL)	119,693	-	-	11,403	-	-
Percentage of AWN Lemmas/BLL	9.7%	12.5%	2.8%	12.5%	3,800.2%	3,787.7%

Table 12 BP Coverage improvement

Figures	Arabic specific characteristic		
	Broken Plurals		
	Original	Extended	Added
No. AWN Synsets	126	1,934	1,434.9%
No. AWN Word-senses	405	2,682	562.2%
No. AWN Distinct Lemmas	120	1,395	1,062.5%
No. Baseline Lexicon Lemmas (BLL)	9,565	-	-
Percentage AWN Lemmas/BLL	1.3%	14.6%	13.3%

The results listed in Table 11 and Table 12 show not only the usefulness of the different AWN extension techniques, but also the significance and the extent of the new content. The most successful outcomes were the addition of the equivalent of 37 thousand times the original number of NE synsets (we created a synset per NE lemma and a one-to-one word-sense), as well as the large number of new word lemmas (15,005 vs. 11,634 in the original version) and new BP forms (1,395 vs. 120 in the original version).

A low coverage improvement was registered for synsets extension (+5.2%). This low increment can be justified as follows: (i) the process used for the automatic extraction of hyponyms was not recursively applied in the current work. Indeed, the hyponyms identified by this process could be used as hypernyms on which we apply the same process again to extract new hyponyms; (ii) the number of extracted snippets was limited to 20 and served as a text collection from which new hyponyms were extracted. Considering a higher number of snippets could increase the number of candidate hyponyms and therefore that of new AWN candidate synsets too. Note that the technique is quite similar to the one used by Snow *et al.* (2005), in that it extends AWN entries with hyponyms on the type level. It does not consider, however, all possible senses for a word type.

With respect to the statistics of the newly proposed AWN release, the previously highlighted gap (see Table 1 in Section 1) relative to the Arabic lexicon (i.e. DIINAR.1) and other WNs considered is now reduced. Table 13 shows the new comparison.

Table 13 Comparison of the extended release of AWN with English WN 3.0 and Spanish WN

Figures	Arabic		Spanish	English
	Original	Extended		
WN Synsets	9,698	10,198	57,424	117,659
WN Word-Senses	18,925	37,463	106,566	206,941
WN Word Lemmas (WL)	11,634	15,005	67,273	155,287
Language Lemmas (LL)	119,693	-	104,000	230,000
Ratio lemmas (WL/LL)	9.7%	12.5%	64.7%	67.5%
Ratio Word-lemmas (WN/English WN)	7.5%	9.7%	43.3%	100.0%
Ratio Synsets (WN/English WN)	8.2%	8.7%	48.8%	100.0%
Ratio Word-senses (WN/English WN)	9.1%	18.1%	51.5%	100.0%

From Table 13, we can see that the extension of AWN now covers around 12.5% of the estimated number of word lemmas in the baseline Arabic lexicon (versus 9.7% without extension). Moreover, after the AWN extension, word-senses represent 18.1% of what already exists in English WN (versus 8.2% before the extension).

Since the resources and techniques used for the proposed AWN extension do not make use of vowelized text, the validation of the new content must be improved by performing Word Sense Disambiguation to introduce the appropriate vowels. The result of this process can help lexicographers to further enhance the accuracy of the extension in a later stage.

For the time being, we have developed a Web interface[16] that presents both the original and the extended content of AWN in order to allow researchers to explore and/or validate the results of the proposed extension. The interface we developed allows:

- Navigating within the AWN hirerachy (synsets tree);
- Consulting the general information of a selected synset (words, part-of-speech, etc.);
- Identifying the source of information (original or extension) using labels (for instance, NS for identifying new synsets, NI for new instances, etc.).

The significance of the new content was also evaluated by conducting new experiments using the AWN-based passage retrieval approach for Arabic Question/Answering, with the aim of showing the impact of AWN extension on performance in this task. The next section recalls the main levels of our approach and the obtained results.

[16] The Web interface can be viewed at: http://sibawayh.emi.ac.ma/awn_extension. The extended release of AWN will also be available after the whole validation process is finished.

3. *Usability of AWN for Query Expansion*

3.1 AWN-based Question Answering

Arabic Q/A is one of the rare cases in which AWN is used as a main resource and where significant experiments are conducted. To give a clear idea about the approach, let us briefly recall that a Q/A system is generally composed of three main modules (Benajiba *et al.* 2007):

(i) *Question analysis and classification module*. In this module a question is analyzed in order to extract its keywords, identify the class of the question and the structure of the expected answer, form the query to be passed on to the PR module, etc.

(ii) *Passage Retrieval* (*PR) module*. This module is one of the most important components of a Q/A system. The quality of the results returned by such a system depends mainly on the quality of the PR module. Indeed, this module uses the query formed by the previous module and extracts a list of passages using an Information Retrieval process (generally a Search Engine such as Google[17] or Yahoo![18]). Thereafter, this module has to perform a ranking process to improve the relevance of the candidate passages according to the user question.

(iii) *Answer Extraction (AE) module*. This module tries to extract the answer from the candidate passages provided by the previous module. In advanced Q/A systems, this module can be designed to formulate the answer from one or many passages.

To our knowledge, there have been just a few attempts meant to build Arabic Q/A systems. Five systems can be mentioned, namely: AQAS (Mohammed *et al.* 1993), QARAB (Hammou *et al.* 2002), ArabiQA (Benajiba *et al.* 2007), QASAL (Brini *et al.* 2009) and AJAS (Kanaan *et al.* 2009). These systems are of limited usefulness, especially, with regard to domains covered, nature of data processed (structured or unstructured), lack of complete experiments with a significant number of questions and/or the number of integrated Q/A modules. Our AWN-based Q/A application aims at overcoming these limitations. Our approach focuses on the PR module since the AE module will succeed in extracting the expected answer only if the PR ranking is relevant. Therefore, our aim is to improve the relevance of the candidate passages generated by this module. Two levels in which AWN has a key role are considered (Abouenour *et al.* 2009b). In the first (keyword-based) level, query expansion (QE) is performed on the basis of semantic relations between synsets (currently limited to synonymy and hyponymy) and the mapping between these synsets and corresponding SUMO concepts. This level tries to improve the recall of the extracted passages with respect to the question keywords and their related terms. The second (structure-based) level refines the relevance of passages by relying on queries that are formed of the question structure and its keywords, together with their related terms. At this stage, the relevance of passages is measured using a Distance Density N-gram model (Buscaldi *et al.* 2010) implemented by a PR tool called JIRS which is also available for Arabic.[19] More details and examples regarding this approach can be found in previous works (Abouenour *et al.* 2009a; Abouenour *et al.* 2009b).

Note that the use of JIRS helps in filtering unvowelized related terms that are generated by the QE process, which relies on the extended AWN content. Hence, the experiments will not be deeply affected by lacks of vowelization in AWN entries.

3.2 Experimental results

Following the experimental process described in our previous work (Abouenour et al. 2009b), new experiments were re-conducted in order to see whether performance of the AWN-based PR approach are improved after extending the content of AWN. It is worth mentioning that this experimental process used well-known Q/A measures (accuracy, MRR and number of correctly answered questions)[20] and that a t-test allowed us to prove the statistical significance of the

[17] http://www.google.com
[18] http://www.yahoo.com
[19] http://sourceforge.net/projects/jirs/
[20] For each question, the accuracy is set to 1 if the correct answer is found in the snippet that is assigned the first rank by the process; otherwise it is set to 0. The question is considered correctly answered if the correct answer figures in one of the first five snippets. The Mean Reciprocal Rank (MRR) is defined as the average

underlying results (Abouenour *et al.* 2010b). More details about the experimental process, as well as questions used, are given in Abouenour et al. (2010b). Table 14 presents the results of the new experiments.

For the sake of comparison, Table 14 also recalls the results that were obtained in Abouenour *et al.* (2010b) with the same 2,224 TREC and CLEF questions (translated into Arabic),[21] without using the AWN-based approach and after using it.

Table 14 Results before and after AWN enrichment

Measures	Without AWN-based approach	Using AWN-based approach				
		Original AWN	After NE extension	After Verb extension	After Noun extension	After whole extended AWN
Accuracy	9.66%	17.49%	25.22%	21.34%	19.21%	26.76%
MRR	3.41	7.98	14.78	13.58	8.55	11.58
Nr. AQ	20.27%	23.15%	35.05%	23.49%	23.89%	35.94%

As we can see, the accuracy, the MRR and the number of correctly answered questions (AQ) were significantly improved after using our approach. Furthermore, the approach exhibited higher performance when it was based on the whole extended content of AWN. Indeed, while the original content allows the application of the approach on 1,470 questions (64.93% of the collection), the extended content raises this number to 1,622 (71.64% of the collection). This brought about an increase in the accuracy from 17.49% to 26.76% (both are higher than the 9.66% registered without the AWN-based PR approach). The MRR also increased from 7.98 to 11.58 and the percentage of answered questions (for which the answer is found in the first five positions) went up from 23.15% to 35.94%. The improvement was also observed when considering each of the CLEF and TREC sub collections separately with the different types of AWN extension. The percentage of questions containing NE keywords is significant (see Table 2), which explains the noticeable performance improvement (35% of answered questions) observed when using the AWN extended with NEs. Thus, the high number of NEs added to AWN synsets helped us to reach this performance.

The increase in performance is not only due to the possibility of applying the AWN-based approach to a higher number of questions, but also to the fact that for each keyword in the question a higher number of related terms are now generated thanks to the extension of AWN. For instance, in the TREC question "من هو الدكتاتور الكوبي الذي أطاح به فيدل كاسترو خارج السلطة في عام 1958 ؟" (Who is the Cuban dictator who was overthrown by Fidel Castro out of power in 1958?), thanks to the AWN extension it was possible to apply the QE process on the verb "أطاح" (overthrown) which was newly added in AWN under the synset ">asoqaTa_v1AR / أسقط". This helped us to get the right answer "باتيستا" (Batista) in the first 10 snippets returned by the Yahoo! API. Applying JIRS on top of this QE process allows drawing this answer to the first 5 snippets considered in our experimental process.

To summarize, within the scope of the experiment just described, we were able to show an improvement in Arabic QA performance using the extended content of AWN instead of the original content. This is a concrete example of the usability of the AWN extension. Nevertheless, the real usability of the extended resource for this specific task (i.e., Arabic QA) remains a subject of future work that will focus on further semantic reasoning based on this resource.

4. Conclusion and Future Works

In the present work, we have focused on the main *coverage* shortcomings in AWN compared to a representative Arabic lexicon and to wordnets in other languages. We have also explained how these shortcomings impact the *usability* of this resource and have been the reasons behind its limited use in Arabic NLP projects. We presented concrete examples of AWN weaknesses and

of the reciprocal ranks of the results for a sample of queries (the reciprocal rank of a query response is the multiplicative inverse of the rank of the correct answer).

[21] Available at: http://www.dsic.upv.es/grupos/nle/downloads.html

evaluated the impact of this resource on Arabic Q/A. Taking this analysis as point of departure, the two-fold aim of the research we reported on was:

- To propose a new release of AWN through the application of semi-automatic extension techniques. Our work allowed us to achieve this aim by means of using, adapting and/or applying existing approaches and resources that were developed for other languages. We succeeded in suggesting new NEs, verbs and nouns (including BP forms) to be added to AWN. We built a new enriched AWN; NEs represent the best content improvement since 433,339 instances were linked to their corresponding AWN synsets. This number is nearly 37 thousand times more than the number of NEs that exists in the current release of AWN. Furthermore, a significant amount of verbs (+122% with respect to the original AWN) was linked to AWN verb synsets. A semi-automatic extraction of noun hyponyms also allowed extracting new AWN synsets and associations. The content of the enriched version of AWN exceeds now the one of the Spanish WN.
- To evaluate the usability of this release in Arabic Q/A. Our evaluation showed that the AWN-based PR module registers higher performance in terms of accuracy (+9.27% improvement), MRR (+3.6 improvement) and number of answered questions (+12.79% improvement) after using the extended AWN.

The present work presents many advantages, particularly considering the fact that it resulted in: (i) the development of AWN by accommodating techniques for its extension and usability, and (ii) a contribution to the work undertaken by the Arabic NLP research community by making available via a Web interface an enriched lexical and semantic resource that can be used in different applications. Future work will focus on enriching AWN with new semantic relations such as meronymy, through the use of pattern discovery techniques, adding new information about verb synsets (such as root variation), building an Arabic YAGO linked to the English one, releasing the extended resource under the same license as the original AWN (CC-by-SA 3.0)**,** conducting experiments to deeply evaluate the usefulness of AWN, and introducing this resource in a semantic reasoning level of the PR module.

Acknowledgements

The work presented in Section 2.2 was done in the framework of the bilateral Spain-Morocco AECID-PCI C/026728/09 research project. The research of the two first authors is done in the framework of the PROGRAMME D'URGENCE project (grant no. 03/2010). The research of the third author is done in the framework of WIQEI IRSES project (grant no. 269180) within the FP 7 Marie Curie People, DIANA-APPLICATIONS - Finding Hidden Knowledge in Texts: Applications (TIN2012-38603-C02-01) research project and VLC/ CAMPUS Microcluster on Multimodal Interaction in Intelligent Systems. We would like to thank Manuel Montes-y-Gomez (INAOE-Puebla, Mexico) and Sandra García-Blasco (Bitsnbrain, Spain) for their feedback on the work presented in Section 2.4. We would like finally to thank Violetta Cavalli-Sforza (Al Akhawayn University in Ifrane, Morocco) for having reviewed the linguistic level of the entire document.

References

Abbès, R., Dichy, J., & Hassoun, M. (2004). The architecture of a standard Arabic lexical database: some figures, ratios and categories from the DIINAR.1 source program. In *Workshop on Computational Approaches to Arabic Script-based Languages*, Coling 2004. Geneva, Switzerland.

Abouenour, L., Bouzoubaa, K., & Rosso, P. (2009a). Structure-based evaluation of an Arabic semantic query expansion using the JIRS passage retrieval system. In *Proceedings of the workshop on computational approaches to Semitic languages, E-ACL-2009*, Athens, Greece, March.

Abouenour, L., Bouzoubaa, K., & Rosso, P. (2009b). Three-level approach for passage retrieval in Arabic question /answering systems. In *Proceedings of the 3rd international conference on Arabic language processing CITALA'09*, Rabat, Morocco, May, 2009.

Abouenour, L., Bouzoubaa, K., & Rosso, P. (2010a). An evaluated semantic query expansion and structure-based approach for enhancing Arabic question/answering. *Special Issue in the International Journal on Information and Communication Technologies/IEEE*. June.

Abouenour, L., Bouzoubaa, K., & Rosso, P. (2010b). Using the YAGO ontology as a resource for the enrichment of named entities in Arabic WordNet. In *Workshop LR & HLT for semitic languages, LREC'10*. Malta. May, 2010.

Ahonen-Myka, H. (2002). Discovery of frequent word sequences in text. In *Proceedings of the ESF exploratory workshop on pattern detection and discovery* (pp. 180-189), London, UK: Springer-Verlag.

Al Khalifa, M., & Rodríguez, H. (2009). Automatically extending NE coverage of Arabic WordNet using Wikipedia. In *Proceedings of the 3rd international conference on Arabic language processing CITALA'09*, May, Rabat, Morocco.

Alotaiby, F., Alkharashi, I., & Foda, S. (2009). Processing large Arabic text corpora: Preliminary analysis and results. In *Proceedings of the second international conference on Arabic language resources and tools* (pp. 78-82), Cairo, Egypt.

Baker, C. F., Fillmore, C. J., & Cronin, B. (2003). The structure of the FrameNet database. *International Journal of Lexicography*, 16(3), 281-296.

Baldwin, T., Pool, P., & Colowick, S. M. (2010). PanLex and LEXTRACT: Translating all words of all languages of the world. In Proceedings of *Coling 2010*, Demonstration Volume, (pp. 37–40), Beijing.

Benajiba, Y., Diab, M., & Rosso, P. (2009). Using language independent and language specific features to enhance Arabic named entity recognition. In *IEEE Transactions on Audio, Speech and Language Processing. Special Issue on Processing Morphologically Rich Languages*, Vol. 17, No. 5, July, 2009.

Benajiba, Y., Rosso, P., & Lyhyaoui, A. (2007). Implementation of the ArabiQA question answering system's components. In *Proceedings of workshop on Arabic natural language processing, 2nd Information Communication Technologies Int. Symposium*, *ICTIS-2007*, April 3-5, Fez, Morocco.

Benoît, S., & Darja, F. (2008). Building a free French WordNet from multilingual resources. *Workshop on Ontolex 2008*, *LREC'08*, June, Marrakech, Morocco.

Black, W., Elkateb, S., Rodriguez, H, Alkhalifa, M., Vossen, P., Pease, A., & Fellbaum, C. (2006). Introducing the Arabic WordNet project. In *Proceedings of the third international WordNet conference*. Sojka, Choi: Fellbaum & Vossen (eds).

Boudelaa, S., & Gaskell, M. G. (2002). A reexamination of the default system for Arabic plurals. Language and Cognitive Processes, 17, 321-343.

Brini, W., Ellouze & M., Hadrich, B. L. (2009a). QASAL : Un système de question-réponse dédié pour les questions factuelles en langue Arabe. In *9th Journées Scientifiques des Jeunes Chercheurs en Génie Electrique et Informatique*, Tunisia.

Brini, W., Trigui, O., Ellouze, M., Mesfar, S., Hadrich, L., & Rosso, P. (2009b). Factoid and definitional Arabic question answering system. In *Post-proceedings of NOOJ-2009*, June 8-10, Tozeur, Tunisia.

Buscaldi, D., Rosso, P., Gómez, J. M., & Sanchis, E. (2010). Answering questions with an n-gram based passage retrieval engine. *Journal of Intelligent Information Systems*, 34(2), 113-134.

Costa, R.P., & Seco, N. (2008). Hyponymy extraction and Web search behavior analysis based on query reformulation. In *Proceedings of the 11th Ibero-American Conference* on AI: Advances *in Artificial Intelligence*, (pp. 1–10).

Denicia-carral, C., Montes-y-Gõmez, M., Villaseñor-pineda, L., & Hernandez, R. G. (2006). A text mining approach for definition question answering. In *Proceedings of the 5th international conference on natural language processing*, *FinTal'2006*, Turku, Finland.

Diab, M. T. (2004). Feasibility of bootstrapping an Arabic Wordnet leveraging parallel corpora and an English Wordnet. In *Proceedings of the Arabic language technologies and resources*, *NEMLAR, Cairo, Egypt*.

El Amine, M. A. (2009). Vers une interface pour l'enrichissement des requêtes en arabe dans un système de recherche d'information. In *Proceedings of the 2nd conférence internationale sur l'informatique et ses applications (CIIA'09),* May 3-4, Saida, Algeria.

Elkateb, S., Black, W., Vossen, P., Farwell, D., Rodríguez, H., Pease, A., & Al khalifa, M. (2006). *Arabic WordNet and the challenges of Arabic*. In *Proceedings of Arabic NLP/MT conference*, London, U.K.

Elghamry, K. (2008). Using the Web in building a corpus-based hypernymy-hyponymy lexicon with hierarchical structure for Arabic. *Faculty of computers and information* (pp. 157-165).

Fellbaum, C. (ed.) (1998). WordNet: *An electronic lexical database*. Massachusetts: MIT Press.

García-Blasco, S., Danger, R., & Rosso, P. (2010). Drug-Drug interaction detection: A new approach based on maximal frequent sequences. *Sociedad Española para el Procesamiento del Lenguaje Natural, SEPLN,* 45, 263-266.

García-Hernández, R. A., Martínez Trinidad, J. F., & Carrasco-ochoa, J. A. (2010). Finding maximal sequential patterns in text document collections and single documents. *Informatica*, 34(1), 93-101.

García-Hernández, R. A. (2007). Algoritmos para el descubrimiento de patrones secuenciales maximales. Ph.D. thesis, *INAOE*. September, Mexico.

Goweder, A., & De Roeck, A. (2001). Assessment of a significant Arabic corpus. In *Proceedings of the Arabic NLP Workshop at ACL/EACL*, (pp. 73–79), Toulouse, France.

Graff, D. (2007). Arabic Gigaword third edition. Linguistic Data Consortium. Philadelphia, USA.

Graff, D., Kong, J., Chen, K., & Maeda, K. (2007). English Gigaword third edition. Linguistic Data Consortium. Philadelphia, USA.

Hammou, B., Abu-salem, H., Lytinen, S., & Evens, M. (2002). QARAB: A question answering system to support the Arabic language. In *Proceedings of the workshop on computational approaches to Semitic languages*, ACL, (pp. 55-65), Philadelphia.

Hearst, M. A. (1992). Automatic acquisition of hyponyms from large text corpora. In Proceedings of the 14th conference on Computational linguistics, COLING '92, Vol. 2 (pp. 539-545).

Kanaan, G., Hammouri, A., Al-Shalabi, R., & Swalha, M. (2009). A new question answering system for the Arabic language. *American Journal of Applied Sciences* 6(4), 797-805.

Kim, H., Chen, S., & Veale, T. (2006). Analogical reasoning with a synergy of HowNet and WordNet. *In Proceedings of GWC'2006, the 3rd global WordNet conference*, January, Cheju, Korea.

Kipper-Schuler, K. (2006). VerbNet: A broad-coverage, comprehensive verb lexicon. Ph.D. Thesis.

Mohammed, F.A., Nasser, K., Harb, H. M. (1993). A knowledge-based Arabic question answering system (AQAS). In *ACM SIGART Bulletin* (pp. 21-33).

Niles, I., & Pease, A. (2001). Towards a standard upper ontology. In *Proceedings of FOIS-2* (pp. 2–9), Ogunquit, Maine.

Niles, I., & Pease, A. (2003). Linking lexicons and ontologies: Mapping WordNet to the Suggested Upper Merged Ontology. In *Proceedings of the 2003 international conference on information and knowledge engineering*, Las Vegas, Nevada.

Ortega-Mendoza, R. M., Villaseñor-pineda, L., & Montes-y-Gõmez, M. (2007). Using lexical patterns to extract hyponyms from the Web. In *Proceedings of the Mexican international conference on artificial intelligence MICAI 2007*. November, Aguascalientes, Mexico. *Lecture Notes in Artificial Intelligence* 4827, Springer.

Palmer, M., P. Kingsbury & D. Gildea. (2005). The Proposition Bank: An annotated corpus of semantic roles. *Computational Linguistics*, 21. USA: MIT Press.

Pantel, P., & Pennacchiotti, M. (2006). Espresso: Leveraging Generic Patterns for Automatically Harvesting Semantic Relations. In *Proceedings of conference on Computational Linguistics Association for computational linguistics*, (pp. 113-120), Sydney, Australia.

Rodriguez, H., Farwell, D., Farreres, J., Bertran, M., Alkhalifa, M., & Martí, A. (2008a). Arabic WordNet: Semi-automatic extensions using Bayesian Inference. In *Proceedings of the the 6th Conference on Language Resources and Evaluation LREC2008*, May, Marrakech, Morocco.

Rodriguez, H., Farwell, D., Farreres, J., Bertran, M., Alkhalifa, M., Mart., M., Black, W., Elkateb, S., Kirk, J., Pease, A., Vossen, P., & Fellbaum, C. (2008b). Arabic WordNet: Current state and future extensions. In *Proceedings of the fourth global WordNet conference*, January 22-25, Szeged, Hungary.

Sharaf, A. M. (2009). The Qur'an annotation for text mining. First year transfer report. School of Computing, Leeds University. December.

Snow, R., Jurafsky, D., & Andrew, Y. N. (2005). Learning syntactic patterns for automatic hypernym discovery. In Lawrence K. In Saul et al. (ed.), *Advances in Neural Information Processing Systems,* 17. Cambridge, MA: MIT Press.

Suchanek, F. M., Kasneci, G., & Weikum, G. (2007). YAGO: a core of semantic knowledge unifying WordNet and Wikipedia. In *Proceedings. of 16th international World Wide Web conference WWW'2007*, (pp. 697-706), May, Banff, Alberta, Canada: ACM Press.

Tjong Kim Sang, E., & Hofmann, K. (2007). Automatic extraction of Dutch hypernym-hyponym pairs. In *Proceedings of CLIN-2006*, Leuven, Belgium.

Toral, A., Munoz, R., & Monachini, M. (2008). Named entity WordNet. In *Proceedings of the Sixth international conference on language resources and evaluation (LREC'08)*, Marrakech, Morocco.

Vossen, P. (ed). (1998). EuroWordNet, a multilingual database with lexical semantic networks. *Kluwer Academic Publishers*, The Netherlands.

Wagner, A. (2005). Learning thematic role relations for lexical semantic nets. PhD. thesis, University of Tübingen, 2005.