Bootstrapping a WordNet for an Arabic Dialect from Other WordNets and Dictionary Resources

Violetta Cavalli-Sforza, Hind Saddiki
School of Science and Engineering
Al Akhawayn University
Ifrane, Morocco
v.cavallisforza@aui.ma, h.saddiki@aui.ma

Karim Bouzoubaa, Lahsen Abouenour
Mohammadia School of Engineering
Mohammed V University-Agdal
Rabat, Morocco
karim.bouzoubaa@emi.ac.ma, abouenour@yahoo.fr

Mohamed Maamouri, Emily Goshey
Linguistic Data Consortium
University of Pennsylvania
Philadelphia, U.S.A.
maamouri@ldc.upenn.edu, egoshey@ldc.upenn.edu

***Abstract*— We describe an experiment in developing a first version of WordNet for Iraqi Arabic starting from Arabic WordNet (for Modern Standard Arabic), Princeton WordNet (for English) and a bidirectional English-Iraqi Arabic dictionary. The resulting initial version of the target WordNet so-constructed was made available to human experts in Iraqi Arabic for correction and evaluation.**

Keywords—language resources; WordNet; automatic creation; bootstrapping; dialect; Arabic

I. Introduction and Motivation

A WordNet (WN) is a lexical database of a given language that focuses on open-class words: nouns, verbs, adjectives, adverbs and adverbials (the latter being mostly nouns, adjectives and participles used in an adverbial role, e.g. 'willingly'). Words belonging to these parts of speech are grouped into cognitive synonyms (synsets), each expressing a distinct concept. Synsets are interlinked by means of conceptual-semantic and lexical relations such as hyponymy and antonymy. The first WordNet was built for the English language (named Princeton WordNet)[1]. The WordNet for Modern Standard Arabic (MSA), followed many years later. The first Arabic WordNet was released in 2007[2] [1][2][3] and followed the development process of English WordNet and Euro WordNet[3] [4]. It utilized the Suggested Upper Merged Ontology[4] as an interlingua to link Arabic WN to previously developed WNs. The most recent version of AWN is AWN 2.0.1 (released in March 2009). To our knowledge no other open source WNs for the Arabic language or its dialects have been developed to date.

While WNs are interesting language resources on their own, allowing the user to explore the relationship of words to each other, they are also useful in a number of language processing tasks requiring an understanding of the meaning of language. Such tasks include information retrieval [5], word sense disambiguation [6], automatic text classification [7], automatic text summarization [8], question answering [9] and machine translation [10], among others (e.g. [11], [12]). However, the construction of a WN requires significant human resources even to obtain a relatively basic subset of the full WN for a language. An alternative to building a WN from scratch is to leverage existing resources, including existing WNs if they exist for similar languages, to provide a "first draft" of a WN resource, which can then be modified and augmented automatically and/or manually by linguists (native speakers of the language).

After the development of the English WordNet and its success when used in the context of many NLP tasks, many languages, including Arabic, went on to build their own WordNets. However, to our knowledge, no WordNet has been developed as a dialect of the WordNet of its corresponding language, and therefore no process and methodology has been proposed and tested for the creation of dialect WordNets.

The work described herein aimed at developing a general process for creation of a basic WordNet for Arabic dialects by using Arabic WN (AWN), Princeton WN (PWN), and an English-Arabic dictionary for a dialect of Arabic. The rationale is as follows. Dialects of Arabic can be presumed to be similar enough in their conceptual organization to Modern Standard Arabic that AWN provides a useful starting point for building a WN for dialects. AWN is linked at the level of synsets to PWN. Using these links, it is possible to use English word information to consult an English-Arabic dialect dictionary. Our project specifically targeted the Iraqi dialect because of the availability of an English-Iraqi bidirectional computational dictionary that was at a more advanced stage of completion than dictionaries for other dialects.

This research was made possible through funding received by LDC, University of Pennsylvania, USA, subcontracted to Al Akhawayn University, Ifrane, Morocco.

[1] http://wordnet.princeton.edu/

[2] http://www.globalwordnet.org/AWN/

[3] http://www.illc.uva.nl/EuroWordNet/

[4] http://www.ontologyportal.org/

The paper is organized as follows. We begin by presenting the resources we used and the pre-processing we needed to perform on them before they could be effectively used. We then explain and justify the automatic process we followed in developing the initial version of the Iraqi Arabic WordNet (IAWN) resource and the raw results obtained through this automatic processing. Successively, we briefly describe the user interfaces we built to browse the resources used, and in particular the tool for exploring and modifying the IAWN. Following that, we present the work performed by human annotators (linguists) and relate some of their reactions to the IAWN resource and associated tools. We conclude with summary of the work performed and some recommendations for future work.

II. Resources Used and Pre-Processing

The construction of the initial IAWN employed AWN 2.0, PWN 2.0 (to which AWN 2.0 is linked), and a computationally accessible Iraqi-English dictionary [13].[5]

A. *Princeton WordNet (PWN)*

Much has been written about PWN [14] and detailed statistics about the contents of different versions are available online.[6] PWN 2.0, though smaller than 3.1, is still much larger than AWN 2.0, containing 115,424 synsets of which most are nouns (79,689), followed by adjectives (18,563), verbs (13,508), and adverbs (3,664). Noun and verb synsets are organized in hierarchies connected by hyponymy/hypernym relationships (a hyponym is more specific than its hypernym) and are also related to other synsets in a variety of ways. Adjective synsets are related to other adjective synsets with similarity links: adjective synsets are similar to adjective satellite synsets and vice versa. Adjective synsets are also connected via specific words to antonym words and via a pertainym link to related nouns. The term ‘adjective satellite’ acquires its meaning from the following: “Two opposite 'head' senses work as binary poles, while 'satellite' synonyms connect to each of the heads via synonymy relations”.[7] Adverbs have an even simpler organization: adverb synsets are not connected to other synsets directly but via their contained words to the adjective from which they derive.

The main information needed from PWN for the IAWN construction task was the synset ID, the part of speech, the English gloss describing and giving examples of use of the concept represented by the synset, the prototype word of the synset (the primary synonym, most representative of the concept), and other synonym words when present.

B. *Arabic WordNet (AWN)*

AWN 2.0 was released in January of 2008; it contains 9,698 synsets, corresponding to 21,813 MSA words, and 6 different link types, totaling 143,715 links. A later version of AWN, 2.0.1, is also available and contains 11,269 synsets, corresponding to 23,841 words, and 22 link types, totaling 161,705 links. Although AWN 2.0.1 is larger than 2.0, correcting some of the problems present in the older version as well as adding synsets and links that simplify some tasks, the choice of the older version does not create serious issues for this work. We return to this subject in the last section of the paper. AWN 2.0 is distributed with a browser application from which it can be downloaded in either XML format and as a set of CSV files.

AWN synsets belong to one of 5 parts of speech: noun (6,438), verb (2,536), adjective (456), adjective satellite (158), and adverb (110). The following are present in AWN 2.0:

- 9,698 ***equivalent*** links connect AWN synsets to their corresponding PWN synsets. There are approximately as many equivalent links as there are AWN synsets, though 10 AWN synsets (currency-related) are equivalenced to 5 PWN synsets.
- 94,841 ***hyponym*** (***of***) links connect PWN hyponym synsets to their PWN hypernym(s) and concern verb and noun synsets only. Many fewer are actually used in AWN because the AWN-PWN equivalence links only involve less than 9% of the PWN links and therefore many of these links are necessarily ignored or collapsed.
- 22,196 ***similar*** links also connect PWN synsets to PWN synsets, occurring strictly between adjective and adjective satellite synsets; they are symmetric, so there are really 11,098 pairs of similar synsets.
- 1,067 ***has_instance*** links connect an AWN synset to another AWN synset representing a named entity.
- 7,993 ***antonym*** links and 7,920 ***pertainym*** links are links between PWN words and not between synsets.

For building IAWN, the most interesting links were the *equivalent* links and *hyponym* links, since they defined the hierarchical structure of AWN indirectly through the PWN *hyponym* links. *Has_instance* links also contribute to the hierarchy but, in AWN 2.0, almost ¾ of these links point to instances that do not actually exist as synsets; of the remaining ones, the target instance synsets were also linked to other synsets via *hyponym* links over ¾ of the time. Since there is no mapping between AWN and PWN at the level of words, *antonym* and *pertainym* links could not really be used. Some of these problems were rectified in AWN 2.0.1. We also did not use *similar* links (although we could have) because they related adjectives only. In AWN, adjective (and adjective satellite) synsets exist, but are not connected to any hierarchy.

C. *Iraqi-English Dictionary (IED)*

We received the IED [13] as a set of distinct XML files for the English-to-Iraqi (E-to-I) directions, containing 10,741 English lexical entries, and Iraqi-to-English (I-to-E) direction, containing 17,457 Iraqi lexical entries . The two directions of the resource both adhere to LMF DTD specifications but, for historical reasons, vary quite a bit in their organization and the information they contain.[8] Moreover, some of the crucial

[5] PWN 2.0 is an older version of PWN; the most recent version is 3.1.

[6] http://wordnet.princeton.edu/wordnet/man2.0/wnstats.7WN.
html, and
http://wordnet.princeton.edu/wordnet/man/wnstats.7WN.html

[7] http://en.wikipedia.org/wiki/WordNet

[8] www.lexicalmarkupframework.org/

information and differences are stored in features and values that are not actually specified in the DTD itself.

The I-to-E direction of the dictionary was the more polished of the two, but our process for creating IAWN (described below), as well as the format and contents of the I-to-E dictionary, made it more convenient to use the IED in the E-to-I direction. To be able to use the IED in both directions, we did some cleaning of the English definitions in the I-to-E direction and augmented the E-to-I part with a part of speech 'hint' based on the contents of the lexical entry (no part of speech was available in this direction of the dictionary). We also converted the XML dictionary into a MySql database to support fast manipulation of the IED data, and added some statistics about content. This served to support lighter and smarter querying during the IAWN creation process in order to manipulate dictionary entries by category and better analyze the situations we needed to treat.

III. IAWN Construction: Process, and Results

Designers following state of the art practices for building WordNets usually start from the CBC (Common Basic Concepts, developed in the context of the EuroWordNet project) and then make changes according to the language at hand. We departed from the common practice by proposing a new method for creating a WordNet resource for dialects of Arabic. We reasoned that, since our target language is a dialect of Arabic and therefore likely to share many of the concepts of Modern Standard Arabic, the final IAWN structure could reasonably be assumed to be quite similar to that of AWN. Therefore, instead of creating the IAWN structure from scratch, it would be simpler to begin with the AWN structure and adjust it manually and/or automatically as needed. Our overall approach, included the following steps:

1. Create the initial IAWN ontology structure to have exactly the same configuration (synsets and semantic links between them) as the AWN ontology structure, but no initial content.
2. Add specific fields for the purpose of tracing how the synsets were automatically filled in or would be modified.
3. Populate the IAWN ontology structurally using the path AWN → PWN → IED → IAWN
4. Analyze the gaps (empty synsets and unused entries in the IED) and manually make the necessary changes and, with the help of linguists, enrichments.

A. *Process*

Thus the first step in our approach was to create an initial linkage structure for the IAWN ontology that paralleled the structure of AWN ontology, at least at the level of hypernym/hyponym links, but without actual content for the synsets. The empty IAWN structure was then filled using information found in AWN, PWN, and the IED. In addition, in order to allow Iraqi linguists and developers to know how synsets were filled, trace information was added to the structure of IAWN.

The actual process of filling the IAWN ontology consisted of three steps, illustrated in Figure 1:

Step 1. Go from each AWN synset to PWN synset(s) mapped to it.

Step 2. Use prototype (and synonym) elements in PWN synset to access the IED in the E-to-I direction.

Step 3. Obtain corresponding Iraqi elements and build the IAWN synset from them, after checking the information against the I-to-E direction.

Step 3 indicates that the processing in Step 2 was in fact somewhat more complex than it appears above. Since the E-to-I direction of the dictionary does not contain part of speech information, it is useful to check a translation obtained from the E-to-I dictionary in the I-to-E in order to make sure that the potential Iraqi translation has the correct part of speech. So we used both directions of the dictionary but differently. Because of the formatting of the English translations in the I-to-E (which sometimes included examples inside the translations themselves), we relied primarily on the E-to-I dictionary to obtain Iraqi words for filling IAWN synsets, and used the I-to-E dictionary to check that we were picking up a translation with the correct part of speech.

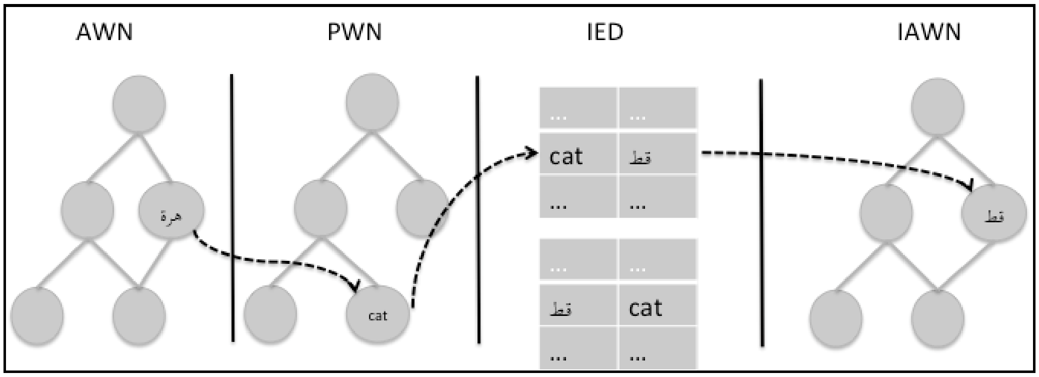


Fig. 1. Visualization of the IAWN filling process: Starting from an AWN synset, follow the link to the corresponding PWN synset, use prototype words and synonyms from PWN to enter the E-to-I IED and obtain Iraqi translations to put in IAWN, filtered by information found in the I-to-E direction.

The following information defines a IAWN synset and is filled in by the automatic tracing process described above:

- **IAWN synset ID**: A unique numerical value that identifies the synset.
- **IAWN prototype word**: Written Arabic and phonetic forms, initially drawn from the IED, making an automatic choice among alternatives, if present.
- **IAWN synonyms**: Computed, similarly to the prototype word, by going through PWN and the IED.
- **Part of speech**: Obtained from AWN, and matched to PWN and IED.
- **Root**: From the I-to-E IED.
- **Word forms**: From the E-to-I IED.
- **Hypernyms and hyponyms**: Initially the same ones as in the corresponding AWN synset.

To provide a trace of how the IAWN synset and its contents were created, the following information is stored:

- AWN to PWN linkage: AWN synset ID, link type between AWN and PWN (currently only equivalent), PWN synset ID
- PWN information: PWN word ID, PWN word, PWN POS, PWN proto (was the PWN word a prototype word for its synset or not)
- PWN to IED linkage: this occurs through prototype words and synonyms
  - English-to-Iraqi: IED LexicalEntry ID, IED senseNumber, IED Definition ID and various other pieces of information about the entry in this direction of the dictionary.
  - Iraqi-to-English: IED LexicalEntryID, IED Sense ID, IED senseNumber. We tried to obtain more information by consulting the dictionary in this direction.

We also keep, for convenience, and to provide as much context as possible for the annotators – for polysemous prototypes and/or sparse synsets, the AWN prototype word, AWN synonyms, and PWN prototype word, though they could be dropped at a later time.

The IAWN filling process has to handle three different cases, due to polysemy of English words and language differences. AWN synsets are linked to PWN synsets representing the corresponding concept, but the link from the PWN synset to the IED is via English words (prototype word and synonyms), which may occur within the IED with different senses and multiple translations for each sense. The cases are as follows:

Case I. In the simplest case, entering the E-to-I IED with the PWN synset prototype word, we find only one sense and a single translation for the corresponding IED LexicalEntry. The Iraqi translation is set as prototype word for the IAWN synset, and translations of PWN synset synonym words are added as IAWN synset synonym words.

Case II. A slightly more complex case is encountered if, entering the E-to-I IED with the PWN synset prototype word, we find multiple translations for the corresponding IED LexicalEntry. The first translation is arbitrarily picked to be the IAWN prototype word and the others are added as synonyms. The PWN synset synonym words are treated as in Case I.

Case III. The most complex case is encountered when, entering the E-to-I IED with the PWN synset prototype word, we find multiple senses for this word (and possibly many translations). In this case an appropriate sense must be heuristically chosen in order to populate the IAWN synset with prototype and synonym words.

A fourth case occurs when there is no mapping from any PWN word to Iraqi through the IED. These four cases jointly give rise to three different statuses for IAWN synsets:

- ***Fairly confident***, including **Cases I** and **II**: The PWN prototype word is not polysemous in the E-to-I IED. This accounts for 73.4% (7882 / 10741) of the lexical entries in the IED dictionary with translations; an additional 48 entries say there is no equivalent translation into Iraqi.
- ***Ambiguous***, representing **Case III**: The PWN prototype word is polysemous: filling the IAWN synset requires care. This case accounts for 26.2% (2805 / 10741) of the lexical entries in the IED with translations; additional 6 entries say there is no equivalent translation into Iraqi.
- ***No mapping***, representing **Case IV**: The IAWN synsets are left empty.

IAWN synsets falling into the ***fairly confident*** case may still not be perfect because the sense found for the PWN prototype word in IED may not be the right sense, and/or because the PWN synonyms may be polysemous, which would add inappropriate translations to the synset. However, the ***ambiguous*** case is the one requiring the most attention to avoid filling the IAWN synset with many inappropriate translations. Two tactics were used for choosing the translations to add to a IAWN synset from IED:

- Tactic 1 follows the basic process specified for the ***fairly confident*** case, but filters the Iraqi translations so that only those with English translations in the I-to-E IED that contain words matching the corresponding PWN synset are retained.
- Tactic 2 is applied on top of the results produced by the first one and is stricter. Fundamentally it tries to pick only those Iraqi translations that are shared by the PWN synset's prototype word and its synonyms, falling back on providing all the translations for a word or a synonym if no overlap with translations of other PWN words was found.

Tactic 2 has greater precision, whereas Tactic 1 has greater recall. However, Tactic 2 does not affect synsets considered

fairly confident, whereas Tactic 1 does. As a result, we chose to apply Tactic 2 to all synsets where it yielded fillers for the IAWN synset, and applied Tactic 1 only to those synsets that could neither be filled by Tactic 2 nor fell in the ***fairly confident*** case.

B. *Results (Gap Analysis)*

The results of the automatic filling process are shown in Table I. We were able to fill 56.2% of the total IAWN synsets with the automatic process; for roughly half of these (25%), we were ***fairly confident*** about the quality of the contents. For 43.8% of IAWN synsets, ***no mapping*** from AWN through PWN and IED could be found. Of the 31.2% of total synsets that were ***ambiguous***, approximately 1/3 has lower quality content because it used Tactic 1 instead of Tactic 2 to gather content. We could improve those numbers a little by manually adding named entity synsets that could not be mapped because there was no routing from AWN to PWN to IED.

The second part of gap analysis concerns how much of the IED was actually used. Results are provided in Table II. The process was able to match and use 3,190 distinct lexical entries out of 10,741, or 29.7% of E-to-I IED, leaving 7,551 entries (approximately 70.3%) unused. As for the other direction of the dictionary, I-to-E, the process was able to match and use 4,203 out of 17,457 (or 24.1%) words in the IED, leaving 13254 unused. There is definitely room for exploring how to make use of the untapped entries and whether there is additional preprocessing of the IED in both directions that can improve use of this resource. We also know that the IED itself has undergone some changes in parallel to this work that will likely affect the performance of our process, so any investigation of how to use the IED resource more fully would need to review these changes, rerun the automatic filling process, and perform again the gap analysis.

The automatically filled synsets were submitted to the attention of Iraqi linguists, through the tools described in the following section, to examine the quality of the results, make changes directly where they could (synset content changes) and suggest modifications where the tools did not afford the ability to make the desired changes directly (changes to the structure of the ontology). The first set of IAWN synsets to be made available for evaluation and modification were the ***fairly confident*** ones. While the linguists were working on those, we developed and refined the process for handling ***ambiguous*** synsets. The linguists' own evaluation of the tools and quality of the automatically generated IAWN resource is presented in Section V below.

TABLE I. RESULTS FOR FILLING OF IAWN SYNSETS

Cases / Confidence	Tactic 2	Tactics 2+1	Total
Easy cases / fairly confident	2426 (25.0%)	2426 (25.0%)	2426 (25.0%)
Ambiguous cases / ambiguous	2363 (24.4%)	2918 (30.1%)	3021 (31.2%)
Empty cases / no mapping found	4909 (50.6%)	4354 (44.9%)	4251 (43.8%)
TOTAL	**9698**	**9698**	**9698**

TABLE II. USE OF IED IN PROCESSING

Cases		**Synset Count**	**Final Average Words per Synset**		**Distinct E-to-I lexical entries matched**	**Distinct I-to-E lexical entries matched**
Fairly confident	**Proto & Syn(s)**[a]	2011	1.6678	1.6999 overall	1,590	1,855
	Syn(s) Only[b]	415	1.8854			
Ambiguous	**Proto & Syn(s)**[a]	2346	2.8500	2.7633 overall	1,600	2,348
	Syn(s) Only[b]	675	2.4622			

[a] Mapped to a PWN prototype & possibly synonym(s)
[b] Mapped to PWN synonyms only

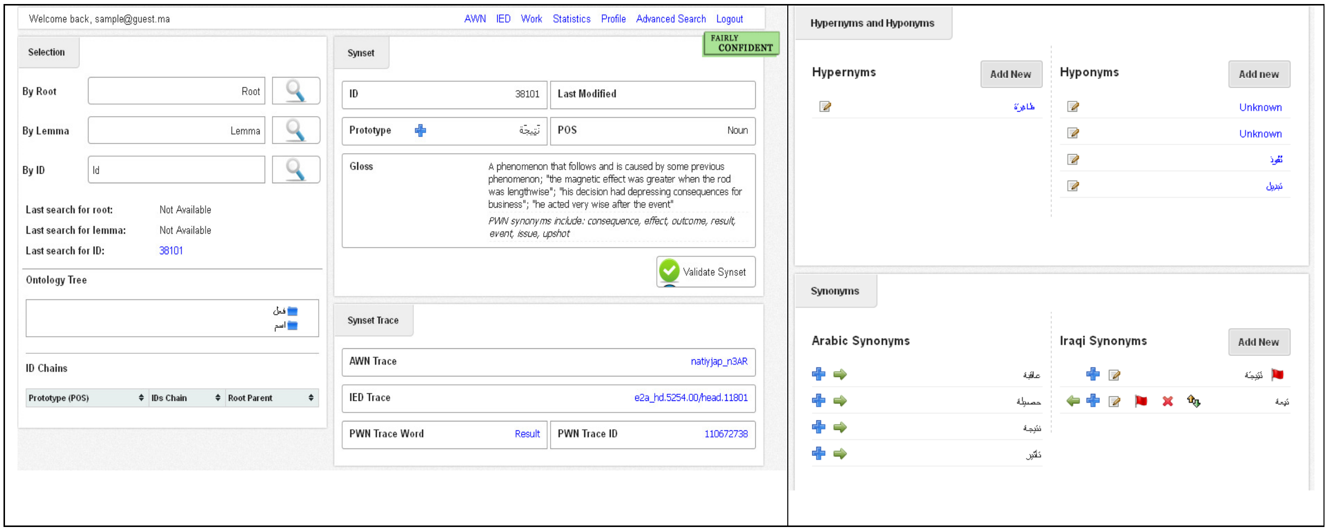
IV. TOOLS FOR BROWSING AND MODIFYING IAWN

The linguists responsible for evaluating the automatically generated IAWN interacted with the resource through a set of three interconnected web-based tools, of which the primary one was the IAWN Interface. For convenience, a web tool for browsing AWN (different from the original Swing browser developed by AWN authors), directly accessible from the IAWN interface was also provided, as well as a tool for browsing the two directions of the IED. In addition, if they desired, they could use the publicly available PWN 2.0 browser[9] and AWN 2.0 Browser[10], which links AWN to PWN synsets in the same display. The AWN and IAWN browser, like the original AWN 2.0 Browser, only shows the hyponym/hypernym structure for nouns and verbs; adjective, adverb and adjective satellite synsets are included, but they are reachable only by direct search by ID, Lemma or Root. The IAWN interface is shown in Figure 2.

The Selection frame is the same as in the AWN tool and can be used to find a synset by Root, Lemma or synset ID. A linguist can view any synset and its details, but can change only synsets s/he is in charge of. It is assumed that mutually exclusive lists of synsets are assigned to different linguists, so no two linguists will be trying to modify the same synset at the same time, but the system enforces this mutual exclusion.

The Synset frame shows the content of a IAWN synset and its status (***fairly confident*** in this example). It is split into several parts. The top of the detailed synset display includes basic information such as synset ID, prototype word, POS, and gloss. The gloss shown comes from the PWN synset to which the AWN synset mirrored by the IAWN synset is linked. If the prototype word has additional information associated with it, it is accessible by clicking the plus icon next to it. Basic synset identifying information is immediately followed by a Synset Trace display area. The trace indicates which AWN and PWN

[9] http://wordnet.princeton.edu/wordnet/download/old-versions/
[10] http://sourceforge.net/projects/awnbrowser/files/awnbrowser/AWNBrowser2.0/



Welcome back, sample@guest.ma

By Root: Root
By Lemma: Lemma
By ID: Id

Last search for root: Not Available
Last search for lemma: Not Available
Last search for ID: 38101

Ontology Tree
فعل
اسم

ID Chains

Prototype (POS)	IDs Chain	Root Parent

ID: 38101
Last Modified
Prototype: نتيجة
POS: Noun
Gloss: A phenomenon that follows and is caused by some previous phenomenon; "the magnetic effect was greater when the rod was lengthwise"; "his decision had depressing consequences for business"; "he acted very wise after the event"

PWN synonyms include: consequence, effect, outcome, result, event, issue, upshot

AWN Trace: natiyjap_n3AR
IED Trace: e2a_hd.5254.00/head.11801
PWN Trace Word: Result
PWN Trace ID: 110672738

Hypernyms
ظاهرة

Hyponyms
Unknown
Unknown
نفوذ
تبديل

Arabic Synonyms
عاقبة
حصيلة
نتيجة
تأثير

Iraqi Synonyms
نتيجة
نتيجة

Fig. 2. Part of the main IAWN browser tool. The Hypernyms and Hyponyms and the Synonyms panes are actually below the browser region at the left. The Additional Comments and Change Requests area would be just below Synonyms but is not shown.

synsets and IED entries were used in creating the IAWN synset. This information can be useful for developers and linguists in understanding the origin of a IAWN node's contents. It can also be used to access information in the AWN and IED tools by clicking on the link, or in other browsers, by copying and pasting. The only action possible here is to validate the synset information, once the user is sure that all changes have been made.

Just below the Synset Trace frame, clickable lists of hypernyms and hyponyms are displayed. The linguist can add to these lists, using the Add New button to search for suitable synsets, but not delete synsets from them or create brand new synsets. This is a precaution to avoid giving the user the power to disconnect a part of the ontology from the remainder by severing a link. Link deletion, and other major structural changes, such as merging or splitting synsets, can be requested via the Additional Comments and Change Requests at the bottom of the screen (not shown). These operations require some careful thinking about the consequences of the change and some work at the database level to preserve the integrity of information in the modified synsets. On the other hand, modifications that are local to the synset and can be easily undone, can be performed directly by the linguist. These include operations such as adding, removing, or otherwise modifying the list of synonyms by changing a word, changing the prototype word (flagged), reordering the synonyms, and accepting a synonym from AWN if it was also used in Iraqi, or rejecting after reconsidering its appropriateness.

V. Linguists' Remarks and Recommendations

Linguists worked with the web-based tools for a period of 71 hours during the month of July 2012 evaluating and modifying the results obtained by the automatic process. According to the linguists' report, the tools worked well from the start and requested fixes or improvements were made quickly, so the tools did not interfere with the linguists' work. They also provided a number of useful comments about the system and the kinds of challenges they encountered.

A. *General Remarks*

Since the project ended in late July, linguists were only able to examine the ***fairly confident*** synsets and made little inroads on the ***ambiguous*** synsets. For those ***fairly confident*** synsets, they remarked that the IAWN synset information provided was largely accurate, with the Iraqi words matching the English glosses and the hypernym/hyponym relationships between synsets based on AWN and PWN carried through successfully to IAWN. However the process of working through synsets systematically led the linguists to provide many new synonyms that either did not make it across from the IED or were not previously recorded. These remarks are indeed reflected in the results shown in Table III, computed from the IAWN frozen on the last day the linguists worked. Remembering that, at final count, there were 2426 synsets in the ***fairly confident*** category, only 7.9% had changes to the prototype word and 16.2% witnessed changes to the number of synonyms in the synset. Changes included removing automatically proposed Iraqi words (6.2%), adding MSA synonyms (23.2%), adding Iraqi words, both prototype (1.6%) and non-prototype (7.3%). No changes were made to synsets for which the automatic process had found no fillers since linguists did not have sufficient time to get to them before the end of the project.

For **ambiguous** synsets, totaling 3,021, the data are very sparse (and we do not know how many synsets the linguists actually looked over). They do, however, show that we can expect a larger proportion of Iraqi synonym words to be inappropriate, as the result of picking up synonyms from the wrong sense of the word in the IED. For both categories, the table shows that, frequently, there are Iraqi words not found by the process or not present in the dictionary being added as prototype and non-prototype words. (We were able to detect

TABLE III. CHANGES INTRODUCED BY LINGUISTS

Annotation/ Change	Affected Synsets		
	Fairly confident	*Ambiguous*	*TOTAL*
Changes to prototype word	192 (7.9%)	32	**224**
Changes to number of synonyms in synset	393 (16.2%)	43	**436**
Iraqi synonym words excluded from their assigned synsets	151 (6.2%)	68	**219**
MSA synonyms included from AWN	563 (23.2%)	57	**620**
Iraqi prototype words added manually	39 (1.6%)	18	**57**
Iraqi non-prototype words added manually	176 (7.3%)	42	**218**

this because words coming from the IED have an Arabic orthographic form and a phonetic form, whereas only the orthographic form could be given to manually added words.)

B. Special Challenges

Referring mostly to work performed on the ***fairly confident*** category, the linguists' report specifies that it was understood that the polysemy of English words could easily lead to the matching of wrong synonyms and that this was not a big problem in general, since the linguists could easily delete a translation that was incorrect for the concept represented by the synset and type in a correct one. However, in cases where words such as 'hit', 'strike', 'walk', 'safety' could have general meanings as well as specific meanings in a sport like baseball, which have no equivalent in Iraqi, the automatic system was picking up general meaning translations from IED and there was no correct translation that could be provided by the Iraqi linguists because of the absence of such concepts in Iraqi. (One wonders indeed if such concepts exist in MSA, so as to cause an AWN synset to be linked to a PWN synset with concepts specific to baseball.)

Another issue reported by linguists is that the translations proposed for concepts that are effectively gerunds in English and verbal nouns for MSA, were getting translated by simple nouns. E.g. وَضْع (waDoE[11]) is linked to "The act of putting something in a certain place or location (PWN synonyms include: placement, location, locating, position, positioning, emplacement)" but the proposed translation is مَوقِع (mawqiE), which just means 'place' or 'location', but is not a verbal noun. In the IED, verbal nouns are placed under their corresponding verbs, therefore, starting from an AWN noun synset and its corresponding PWN, the process would need to parse the PWN definition in order to know that it was a noun representing "the act of Xing", recover the verb "to X" in the E-to-I IED and search for the verbal noun therein. Searching in the I-to-E IED for the English word would present even more problems, as it may not even be present as a translation of the Iraqi term sought.

Concerning the general ability to make structural changes to the IAWN ontology, during the first few sessions, the linguists typed some of their recommended changes into the "Additional Comments and Change Requests" text box within the tool, but as time went on and the linguists gained the ability to make more changes directly because of tool improvements, they stopped using this part of the tool entirely. (We found, in fact, only 12 suggestions.) The linguists found that the tool allowed them to directly make any changes in which they felt confident. As for the more complicated changes in structure, it was most often the case that the problem could be solved in several ways, each of which would have different implications for IAWN. For these reasons, the linguists preferred to take their own notes and use such examples as material for analysis and discussion about the data rather than recommending any specific remedies. One remark did, however, come up with respect to structural changes. Just as English makes some distinctions that are not necessarily meaningful in Iraqi (or even MSA), a dialect like Iraqi may make distinctions that are not meaningful in English and therefore are not reflected in PWN. The linguists provided the example of the word حمامة (HamAmap) 'pigeon'. Since Iraqis keep them as pets, train them to fly competitively, and even buy and sell them for income, if IAWN were to accurately reflect the colorful Iraqi vocabulary on this subject, there would be various hypernyms, hyponyms, and synsets all related merely to pigeons. In this case it would be desirable to be able to add brand new synsets, which do not necessarily link to either AWN or PWN synsets, something which the IAWN tool does not yet allow.

C. Additional Recommendations by Linguists

The linguists who worked on evaluating and modifying the results of the automatic process strongly recommended that, in determining what should be the final content of any Iraqi Arabic Word Net, it would be essential to consult Iraqis from a variety of ages and backgrounds. Even though the focus of this study was on the lexicon of Baghdadi variety of Iraqi, individuals of different gender, class, generation and level of education would surely bring different knowledge of different vocabulary. A case in point is the use of MSA synonyms in IAWN. We included them in the interface so that, if appropriate, the linguist could just copy them over one at a time from the AWN synset to the IAWN one, under the assumption that dialects often do use words from the Standard Arabic. The data in Table III indeed show that this was a good assumption and a useful feature. However, linguists were not unanimous about this matter. The two Iraqi linguists included one male and one female; the linguists themselves remarked that the male linguist tended to exclude many MSA synonyms that the female linguist would identify as also being legitimate Iraqi words, though the gender association could have been purely coincidental.

A final additional request was to allow linguists to modify the PWN glosses if the PWN information were going to remain part of the IAWN resource, since it sometimes gave

[11] Buckwalter transliteration is used.

glosses that were inappropriate to the Iraqi culture, e.g. defining 'tribe' as "A social division of (usually preliterate) people".

VI. Summary and Future Work

We have described an approach to building a WordNet of a dialect of Arabic, Iraqi, that leverages the existence of a WN for Modern Standard Arabic, linked to Princeton WN, and a bidirectional English-Iraqi dictionary. Instead of following the conventional approach for building WordNets, which builds a resource from scratch starting from a core of Common Basic Concepts and adds content manually, we have 'bootstrapped' Iraqi Arabic WN from the available resources and submitted it to the attention of native speakers of Iraqi Arabic, who evaluated some parts of the results obtained and provided their feedback and recommendations.

Our approach started with the assumption that a WN for a dialect of a language would have a structure not too different from that of the base language. Therefore we started with an empty WN shell for Iraqi that reflected the structure of Arabic WN. An automatic process then used the synset-to-synset links between Arabic WN and Princeton WN and the English words in Princeton WN synsets to find Iraqi translations in the English-Iraqi and Iraqi-English dictionary. These words were added to the Iraqi WN synsets. Because of project deadlines, the work performed by linguists was largely limited to the better quality portion of IAWN, the one whose contents could be generated with a certain confidence because the English words that serve as the link between Princeton WN and the English-Iraqi dictionary have unique meanings in the dictionary. For those synsets, linguists effected relatively few changes, and generally gave a positive evaluation for the tool. The linguists' work did not get far enough on synsets whose contents were obtained using words with multiple meanings. We obviously expect the quality of those synsets to be lower, but the process of cleaning up the synset makes deletion of inappropriate translations a one-click action.

This initial experiment provided useful information for future work. In the first place, any such approach is as good as the resources it is based on. In that respect, some quality can be gained by starting with a more recent version of Arabic WN (2.0.1) that corrects some of the problems found in the version we used (2.0), and enriching that resources for the benefit not only of Iraqi but of other dialects of Arabic as well. Secondly, we know that the automatic process ended up using 29.7% of English-to-Iraqi and 24.1% of the Iraqi-to-English dictionary, so any further improvements to the automatic process should start with an analysis of what is not being used, why it is not being used and how/whether it could be used to allow the automatic process to fill more synsets. Thirdly, we know that the dictionary itself has undergone some improvements since the version used for this experiment, so the newer version of the dictionary itself should be used. Additional processing on contents of Princeton WN, as well as on the two directions of the English-Iraqi dictionary will probably help make the content of those resources yield more and higher quality results. Finally, since no degree of automatic processing is likely to yield as high quality a resource as one benefiting from human evaluation and modification, we can improve the tools used by linguists to interact with the WN resource to give more flexibility in performing modifications. We should not forget the linguists' own advice: the final resource should reflect the opinions of people from a variety of backgrounds, since each will have something to contribute.

Acknowledgment

Special thanks go to Ayoub Khobalatte and Mariam El Mezouar, of Al Akhawayn University, for all their work on the AWN, IAWN and IED web interfaces to the resources. We are also indebted to David Graff, of the Linguistic Data Consortium for his assistance in checking the installation of databases and software and performing quality control on the conversion of IAWN to Kyoto LMF format.

References

[1] W. Black et al., "Introducing the Arabic WordNet project," in Proceedings of the Third International WordNet Conference, Fellbaum and Vossen (eds), 2006.

[2] S. Elkateb et al., "Building a WordNet for Arabic," in Proceedings of the Fifth International Conference on Language Resources and Evaluation, Genoa, Italy, 2006.

[3] H. Rodriguez et al., "Arabic WordNet: current state and future extensions," in Proceedings of The Fourth Global WordNet Conference, Szeged, Hungary, January 22-25, 2008.

[4] P. Vossen, (ed.). EuroWordNet: A Multilingual Database with Lexical Semantic Networks, Dordrecht: Kluwer Academic Publishers, 1998.

[5] M. Rila, T. Tokunaga, and H. Tanaka, "The use of WordNet in information retrieval," in Proceedings of Workshop on Usage of WordNet in Natural Language Processing Systems, 1998.

[6] R. Navigli, "Word sense disambiguation: a survey," ACM Computing Surveys, Vol. 41, No. 2, pp. 1–69, 2009.

[7] Z. Elberrichi, A. Rahmoun, and M.A. Bentaalah, "Using WordNet for text categorization," in The International Arab Journal of Information Technology, Vol. 5, No. 1, January 2008.

[8] C. Dang, X. Luo, "WordNet-based document summarization," in Proceedings of Seventh WSEAS Int. Conf. on Applied Computer & Applied Computational Science (ACACOS '08), Hangzhou, China, April 6-8, 2008.

[9] P. Clark, C. Fellbaum, and J. Hobbs, "Using and extending WordNet to support question-answering," in Proceedings of the Fourth Global WordNet Conference, University of Szeged, Hungary, pp. 111–119, 2008.

[10] K. M. Anwarus Salam, M. Khan and T. Nishino. "Example based English-Bengali machine translation using WordNet," in Proceedings of Triangle Symposium on Advanced ICT (TriSAI), Tokyo, 2009.

[11] H. Kim, S. Chen, and T. Veale, Analogical reasoning with a synergy of HowNet and WordNet," in Proceedings of the Third Global WordNet Conference (GWC'2006), Cheju, Korea, January 2006.

[12] A. Wagner, "Learning thematic role relations for lexical semantic nets," PhD. thesis, University of Tuebingen, 2005.

[13] M. Maamouri (ed.), Iraqi/English dictionary database, v1.1, Linguistic Data Consortium, January 2012. [LDC catalog ID: LDC2012R21].

[14] C. Fellbaum (ed.), WordNet: An electronic lexical database. Cambridge, MA: MIT Press, 1998.