A Prototype English-Arabic Dictionary Based on WordNet

William J. Black and Sabri El-Kateb

UMIST, Department of Computation, Manchester, M60 1QD, UK
Email: wjb@co.umist.ac.uk, Sabri.El-Kateb-2@student.umist.ac.uk

Abstract. We report on the design and partial implementation of a bilingual English-Arabic dictionary based on WordNet. A relational database is employed to store the lexical and conceptual relations, giving the database extensibility in either language. The data model is extended beyond an Arabic replication of the word↔sense relation to include the morphological roots and patterns of Arabic. The editing interface also deals with Arabic script (without requiring a localized operating system).

1 Introduction

Our goal is the development of an expandable computer-based lexical and terminological resource to aid the working translator or information scientist working with technical terminology in Arabic. [3] The plan has been to use a relational database representation of the Wordnet as a backbone on which to hang translation equivalents and information about domain-specific technical terminology. We are therefore concerned with the potential for the WordNet data model to be extensible. Accounts of earlier versions of the design are given in [2,1]. The present paper gives an up-to-date picture of the data model and design, together with information on implementation and on the lexicographer's user interface.

The EuroWordNet [7,8] approach to multilingual resource development has emphasized the separate integrity of the dictionaries in the different languages, and provided an additional bilingual index to support the search for translations. The effort reported here is on an altogether more limited scale, and stores the data for the different languages in the tables of a single database. In keeping with this small scale, the bilingual dictionary does not currently maintain either glosses or examples in the second language, although there is nothing to prevent the data model being so augmented in the future.

When considering languages more closely related to English, developing a multilingual wordnet can be as simple as providing the mapping of foreign words to synsets. Arabic has an extensive system of derivational morphology that embodies important semantic relations, which ought to be reflected in any conceptual dictionary. The prototype dictionary described here embodies these kinds of lexical relation as well as those present in the WordNet. It also supports Arabic script rather than relying on a transliteration.

The remaining sections discuss Arabic morphology; the data model used and its practical realization in a DBMS; the encoding of Arabic morphological information; the facilities of the current user interface for editing and updating the data; how lexical mismatches are handled.

Petr Sojka, Karel Pala, Pavel Smrž, Christiane Fellbaum, Piek Vossen (Eds.): GWC 2004, Proceedings, pp. 67–74.

2 Arabic Morphology

Arabic morphology is described as "non-concatenative", not because of any absence of prefixes and suffixes, but because affixation is not the only morphological process supporting inflection and derivation.

Arabic [4] has a word structure whereby related forms share a sequence of three or four consonants, following each of which are different vowels, according to the form. That is, words have a basic structure CVCVCV or CVCVCVCV. Prefixes and suffixes also contribute to the differentiation of forms. There are only three distinct vowels /a/, /ι/ and /u/, but these also come in long variants, indicated in transliterations by a following colon.

2.1 Arabic Script

Mst ltrt nglsh spkrs cn dcd txt n whch thr r nly cnsnts, thanks to the redundancy in the script. Arabic readers do this all the time, because most vowels are suppressed from the written language, including dictionary citation forms. The vowels can be indicated by diacritics placed above or below the consonant that precedes them, when necessary for expository purposes.

In addition to the three vowels, there are 25 consonants in the script, and as Arabic is a cursive script, the letters take different forms according to whether they occur in initial, medial or final position in the written word.

Table 1 illustrates the way that semantically related forms are derived from a common root, with a set of words sharing the consonant sequence /w/ /l/ /d/. (The Arabic script letters for these consonants are و, ل and د respectively.)

Table 1. Words derived from a common root

Word	Translit.	Pattern	Pattern translit.	English
ولَادة	wila:dah	فِعَالَه	fi'a:lah	delivery
تولِيد	tawli:d	تَفعِيل	taf'i:l	generation
توَالد	tawa:lud	تَفَاعُل	tafa:'ul	reproduction
وَالد	wa:lid	فَاعِل	fa:'il	male parent
مولُود	mawlu:d	مَفعُول	maf'u:l	new born baby
مولد	mawlid	مَفعِل	maf'il	birth

2.2 Inflection and Derivation

The same kinds of word change are used to inflect as well as derive forms in Arabic. Inflected forms do not customarily occur in printed dictionaries, and are therefore not of interest to the dictionary compiler. Whilst an on-line dictionary like the WordNet can allow users to enter queries with inflected forms, if there is a morphological analyser or lemmatizer component, dictionary users know that it is the base or citation form they should expect to use.

Derivational morphology is another matter. In conventional dictionaries, it is customary for some derived forms to be made completely subsidiary to the headword, rather than having a separate entry. In WordNet 2.0, derivational relations between nouns and verbs can be traced, and these relations ought to be traceable in any other dictionary based on conceptual principles. Arabic dictionaries (mono- or bi-lingual) are sometimes ordered according to morphological roots, with large numbers of forms (possibly out of alphabetic sequence) being listed subsidiary to them.

In Arabic, speakers are much more conscious of derivational morphology, since the bulk of the vocabulary has a systematically encoded derivation from a few thousand roots (which are all verbs). In table 1, we see for example, that the vowels in the word transliterated as `wa:lid` are a long /a:/, an /ι/ and a null vowel. Words with different roots share this pattern, which has been transliterated `fa:'il`.[1] Seeing the words that share a pattern, one can be tempted to try to encode the meaning of the form as a semantic feature. However, such features are difficult to encode and not always productive.

Derivation and Borrowings The process of derivation has proved to be flexible enough to derive from non-native words. Arab linguists stress the need to make borrowed terms concordant with the phonological and morphological structure of Arabic, to allow acceptable derivatives. For example, the English term *oxide* is pronounced `oksa:yid` in Arabic but it is modified to `uksi:d` in order to generate the derivatives shown in table 2.

Table 2. Derivations from a borrowed word

Arabic Word	pattern	English Word
aksada	fa'lala	oxidize
muaksad	mufa'lal	oxidized
aksadah	fa'lalah	oxidation
taaksud	tafa'lul	oxidation

Morphology in the Bilingual Wordnet We conclude that in an Arabic-English bilingual wordnet, the derivational root and form of each content word should be stored, since this way of semantically linking words is a basic expectation of a literate Arabic speaker. However, it is not considered appropriate to attempt to 'decode' the patterns as semantic features or named relations.

[1] All patterns are written by convention with the same consonants /f/ /'/ and /l/ (and short vowels are written as diacritics). Textbooks often refer to the patterns by number or mnemonic rather than using these consonants as a skeleton.

3 Strategy for Building the Arabic-English Wordnet

One way to construct a bilingual wordnet would be to write lexicographers' files and compile a database with the grinder. However, the data for the English and Euro WordNets are available in alternative formats, including XML and Prolog. Persistently stored in a relational database, the data can be readily extended or modified in real time without a compilation step. New tables have been constructed to encode translations between synsets and Arabic words, roots and patterns.

We used Prolog clauses, edited to turn them into database tables via the comma-separated file format, as described in [2]. For efficient hyponymy navigation, we store with each synset, the path to it from the top of the tree and all its immediate hyponyms. On-demand selective tree display is acceptably fast.

3.1 Adding Data for Other Languages

There are several alternative ways to add a second and subsequent language to a sense enumerative lexicon [9], who discuss ways to link the senses in separate language-specific conceptual lexicons. It is equally possible to extend the data model to create a single multilingual repository. In our design, there is a single set of conceptual relations shared by the two (or more) languages. To make the database multilingual, the basic need is to provide the word↔sense table[2] for the additional language(s). Three possible extensions to the data model are:

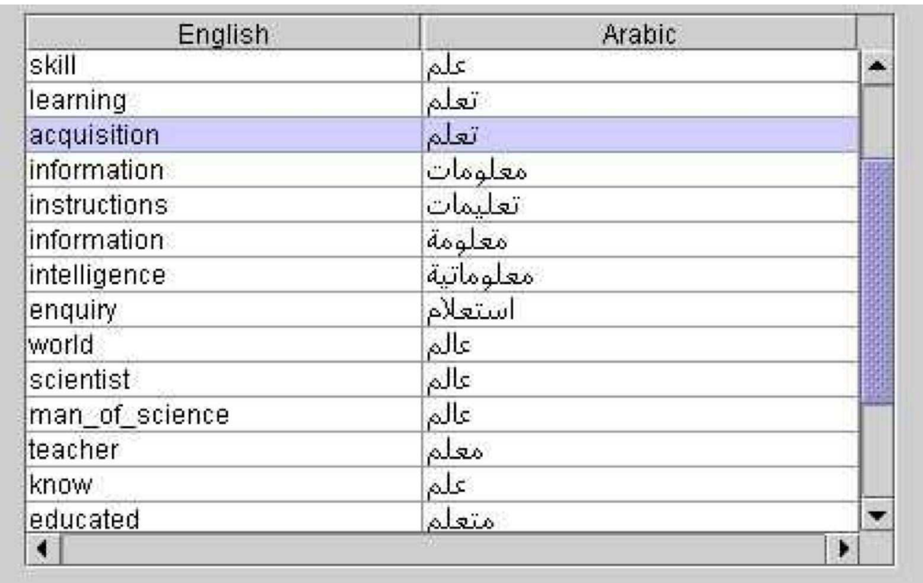
1. Label the *word* column *English*, and add columns for each language.
2. Add a column encoding the language of the table row.
3. Reproduce a word↔sense table for each language.

Alternative (i) is not very attractive, as it implies a change to the database structure whenever an additional language is added to the database, although it is reasonably space-efficient if most words have equivalents in the various languages. Between alternatives (b) and (c), although the former is the more language-independent, we actually adopted the latter despite the language identity's embodiment in the table name. This was because of additional columns of attributes (described below) needed for Arabic, but not for other languages.

4 Words, Roots and Patterns in the WN_S_ARABIC Table

The Arabic equivalent of the WN_S table has the root and pattern of each word as additional columns. This allows the system to support queries based on words, roots or patterns, as well as via synonymy, hyponymy and the other Wordnet relations, and by English translation. Figure 1 shows the result of a query based on a shared root with the query word. In the database as presently constituted, words are written as cited in conventional dictionaries, without diactritics, although patterns are, of necessity, written with diacritics.

[2] This table has attributes synset_id, word, part of speech, and integers indicating the relative frequency of word within synset and of the sense of the word. A join of the table with itself finds either the synonyms of a word or its alternative senses.



English	Arabic
skill	علم
learning	تعلم
acquisition	تعلم
information	معلومات
instructions	تعليمات
information	معلومة
intelligence	معلوماتية
enquiry	استعلام
world	عالم
scientist	عالم
man_of_science	عالم
teacher	معلم
know	علم
educated	متعلم

Fig. 1. Query result with derivationally related Arabic words

With a morphological analyzer, it should be possible to dispense with the *word* column in the database, deriving it on demand from the root-pattern combination, and also to provide the diacritic form and/or transliterations for the benefit of learners of Arabic.

5 Editing Functionality and the User Interface

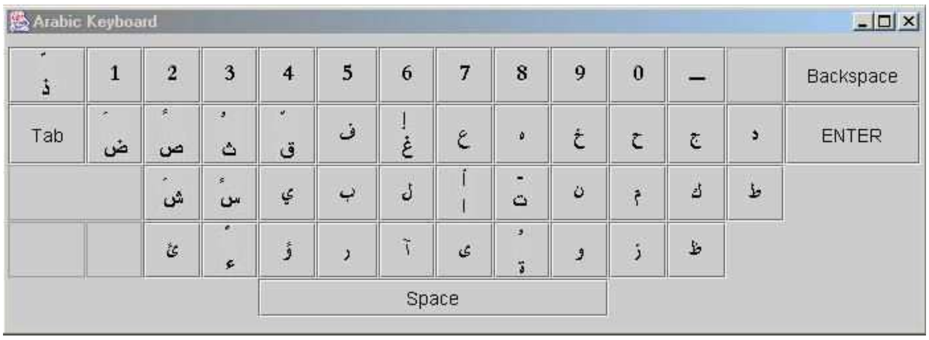
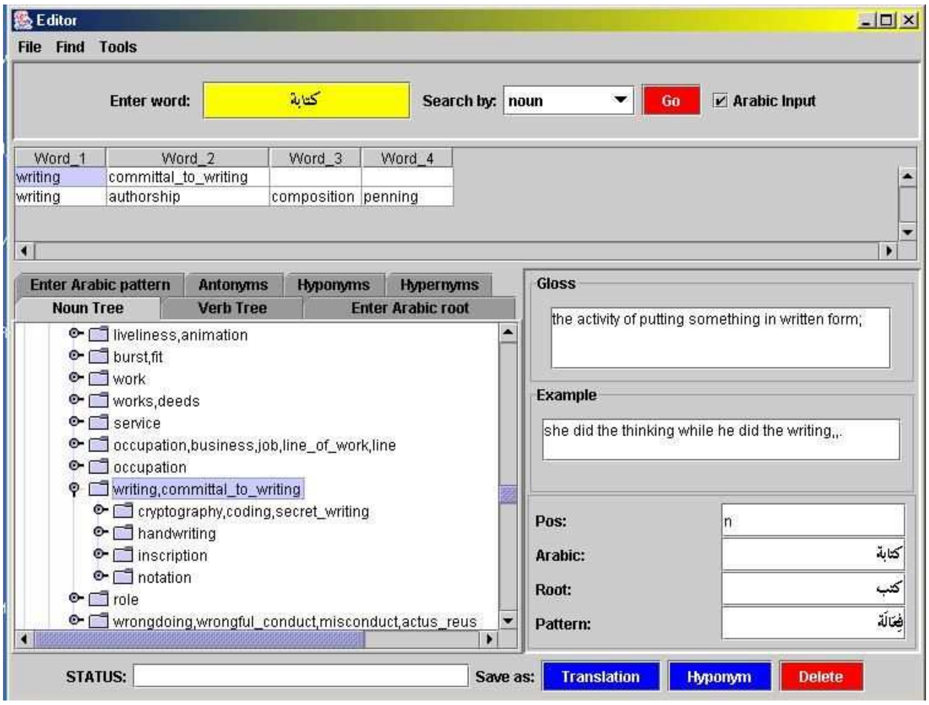

Arabic Keyboard
ذ 1 2 3 4 5 6 7 8 9 0 — Backspace
Tab ض ص ث ق ف غ ع ه خ ح ج د ENTER
ش س ي ب ل ا ت ن م ك ط
ئ ء ؤ ر آ ى ة و ز ظ
Space

Fig. 2. Simulated Arabic keyboard

Users and editors of a wordnet have different needs. A read-only interface can use formatted displays of synset lists, hyponymy trees etc. For an editor, there has also to be the possibility of making a single word or sense from those retrieved or browsed *current*. Overall, the editor must support similar user operations to the EuroWordNet Polaris editor [9]. New items added to the database are then linked into sense relations like hyponymy, relative to the *current* synset. The information displays treat each element as a distinct object rather than as text. Figure 3 shows the current version of the interface and examples of the controls necessary to support updating. All updates are made relative to an item previously retrieved, so the interface has a query facility (the top panel in figure 3). This allows words to be entered in either English or Arabic (and additionally Arabic roots and patterns), and a number of alternative queries invoked (via the pull-down menu). Since words typically have multiple senses, the initial response to a query is to display a word↔sense matrix, as a table that allows cells, rows or columns to be selected (shown in the upper part of figure 3). Selecting a cell or a row makes a particular synset current. This in turn enables the tree-view to be generated and focused around the selected sense. At the same time, the gloss and examples (if any) for the selected sense are also retrieved and displayed. Any updates are made relative to the synset currently shown as selected.



Enter word: كتابة
Search by: noun
☑ Arabic Input

Word_1	Word_2	Word_3	Word_4
writing	committal_to_writing		
writing	authorship	composition	penning

- liveliness,animation
- burst,fit
- work
- works,deeds
- service
- occupation,business,job,line_of_work,line
- occupation
- writing,committal_to_writing
  - cryptography,coding,secret_writing
  - handwriting
  - inscription
  - notation
- role
- wrongdoing,wrongful_conduct,misconduct,actus_reus

Gloss: the activity of putting something in written form;

Example: she did the thinking while he did the writing,,.

Pos: n

Arabic: كتابة

Root: كتب

Pattern: فِعَالة

STATUS:

Fig. 3. Editor user's interface

Updates are confined to the entry of Arabic words equivalent to or related to the selected displayed synset. The editor enters the corresponding Arabic word, root and pattern in the fields in the panel towards the bottom right of Figure 3, pressing the button labeled

“Translation” to save the new word’s details. This creates an entry in the WN_S_ARABIC table, with the same synset as the current one. Deletions from that table can be accomplished after retrieval of the item directly or via its English translation sysnet becoming current during browsing.

When a Direct Translation is not Possible There are numerous well-known conceptual difficulties in translating between languages. Both English and Arabic have many vocabulary items with no direct equivalent in the other language. Some of the fields in which these occur are religion, politics, food, clothing, etc. A small selection of Arabic words, all to do with Ramadan, and with no direct English equivalent is given in table 3.

Table 3. Words derived from a common root

Word	Transliteration	Meaning
سحور	suhu:r	light meal taken before starting a new day of Ramadan
مسحرَاتي	musahara:ti	man who beats a drum in the streets (before dawn) to wake people up to eat before they start a new day of fasting
إفطَار	ifta:r	meal at the end of daily fasting during Ramadan
مدفع افطَار	midfa’ ifta:r	gun announcing the end of daily fasting during Ramadan
عمرة	umra	visit to the holy shrines in Mecca and Madina out of the time of the Pilgrimage

Where a word-root-pattern is entered having no English translation, a new Synset_id is allocated. Then this must be linked to its nearest hypernym (by adding a new row to the English table), and a new row to the Arabic version of the word↔sense table. An English gloss should also be added. What the user has to do in such a case is to find a suitable hypernym by search or browsing, prior to pressing the (save as) Hyponym button.

6 Conclusions and Further Work

We have described the design and partial implementation of a bilingual WordNet-based resource for English and Arabic, supported by a software framework built round a relational database. This enables us to store interesting conceptual relations additional to those in the original WordNet, and for the database to be extensible, particularly in the second language. To support the needs of end users, we will also need to incorporate a treatment of morphology. The original plan had been to adaopt the implementation by Ramsay and Mansur [5], although we are actively seeking alternatives that do not require multiple computer languages in the implementation. Other end-user-oriented features will be to widen the types of query supported, including free text queries of the glossary and example entries [6]. As computational linguists working on text mining applications, we are keen to experiment with the indirect use of the Arabic lexicon in revealing semantic relations useful to tasks such as WSD.

References

1. Black, W. J., El-Kateb, S.: Towards the design of an English-Arabic termonological and lexical knowledge base. In Proceedings of the Workshop on Arabic Natural Language Processing, ACL-2001, Toulouse, France (2001).
2. Denness, S. M.: A Design of a Structure for a Multilingual Conceptual Dictionary. MSc dissertation, UMIST, Manchester, UK (1996).
3. El-Kateb, Sabri: Translating Scientific and Technical Information from English into Arabic, MSc dissertation, University of Salford, UK (1991).
4. Holes, C. Modern Arabic. London: Longman (1995).
5. Ramsay, A. and H. Mansur.: Arabic Morphology: a categorial approach. Proceedings of the ACL2001 Workshop on Arabic Language Processing: Status and Prospects, Toulouse, July 6th, 2001.
6. Sierra, G. and McNaught, J.: Design of an onomasiological search system: A concept-oriented tool for terminology. Terminology **6**(1), 1–34. (2000).
7. Vossen, P.: Introduction to EuroWordNet. In: Nancy Ide, DanielGreenstein, Piek Vossen (eds), Special Issue on EuroWordNet. Computers and the Humanities, **32** (2–3) (1998), 73–89.
8. Vossen, P.: EuroWordNet: A Multilingual Database with Lexical SemanticNetworks, Dordrecht: Kluwer Academic Publishers (1998).
9. Vossen, P., Dez-Orzas, P., Peters, W.: The Multilingual Design of EuroWordNet. In: P. Vossen, N. Calzolari, G.Adriaens, A. Sanfilippo, Y. Wilks (eds.) Proceedings of the ACL/EACL-97 workshop Automatic Information Extraction and Building of Lexical Semantic Resources for NLP Applications, Madrid, July 12th, 1997.