English-Arabic Dictionary for Translators

Sabri Elkateb and Bill Black
Department of Computation
UMIST,PO Box 88, Manchester M60 1QD
Sabri.El-Kateb2@student.umist.ac.uk , wjb@co.umist.ac.uk

Abstract

We present a design of a computerized bilingual Arabic-English-Arabic conceptual dictionary for translators. This study is an attempt to develop a structure whose query mechanism is largely based on the query process implemented in WordNet, the Princeton Lexical reference database, in the form of a conceptual dictionary (Miller, 1990), (Beckwith and Miller, 1990). Our goal is not only to add the Arabic language to the present database, but also proposing some important features in an attempt to enhance the value of the design. Our design will provide additional search facilities, like syntagmatic and paradigmatic relations between different parts of speech as well as roots, patterns and derivatives of words.The editing interface also deals with Arabic script (without requiring a localized operating system).

1 Introduction

The notion of what a dictionary is has undergone a dramatic change with developments in computational lexicography and in computational linguistics. A declarative representation of word and sense relations makes possible adhoc queries with which the user (or the natural language processing system) can find syntactic, conceptual, morphological, phonetic information about a word, and its possible translations in other languages. Equally, declarative representations used in current lexical and terminological knowledge bases can enable the search for words realizing a concept, sense or lexeme, e.g. as proposed in (Sierra and McNaught, 2000). We describe the conceptual design of a terminology base, based on a 'backbone' derived from a relational model of the WordNet (Denness, 1996). The data model is extended beyond an Arabic replication of the word sense relation to include the morphological roots and patterns of Arabic.

2 Approach

We mainly aim at developing an expandable, browsable and searchable computer-based lexical and terminological resource for translators and information scientists working with technical terminology in Arabic. Besides the desire that this dictionary can meet the needs of various groups of users, it is mainly intended for Arab translators who seek to have satisfactory information about a word and an adequate representation of its form, structure and senses.

One of the interesting modes of organisation of this conceptual dictionary is that indexing of the sets of words replaces alphabetical order. The set of words or synonym sets known in this implementation as the 'synset' represents a concept. A word - concept relation supports three query types with both word and concept indexed:

- Senses of word
- Words expressing a concept
- Synonyms of a word.

3 WordNet Model

WordNet is a monolingual English Language on line lexical resource developed at Princeton University by psychology professor George Miller. This lexicon is organised in terms of word meaning rather than word forms. WordNet organises the lexicon by semantic relations on the basis of synonymy. Synonymy is a semantic relation between two words with different forms and similar meanings. Table 1, extracted fromthe distribution of the WordNet in Prolog form, and edited table format, shows how this may be viewed in tabular form. Wordnet represents senses as the collection of words having that sense - a set of synonyms, or a synset. The

sense is no more than that set of words that denote it, but in a database it is convenient to represent each such set with a unique identifier, shown in the table as Synset_No.

Synst_No	Word#	Word	Cat	S#
100001742	1	entity	n	i
100003135	1	organism	n	1
103447508	1	plant	n	1
105054818	1	plant	n	2
106962451	2	flora	n	1
201241292	1	plant	v	1

Table 1: Word-sense relations derived from WordNet

4 EuroWordNet Model

EuroWordNet is a multilingual database with various wordnets for several European languages (Dutch, Italian, Spanish, German, French, Czech and Estonian). The wordnets adopt the same structure implemented in American wordnet for English (Princeton WordNet (Miller, 1990)). There is a unique language-internal system of lexicalizations for each participant wordnet, and each wordnet is linked to an Inter-Lingual-Index or ILI, based on the Princeton wordnet. This index makes the languages interconnected, i.e. search can go from the words in one language to similar words in any other language. EuroWordNet approach aims at building the wordnets mainly from existing resources. Each site in the project can build their language-specific wordnet using their tools and resources available in previous national and international projects.

5 Word-sense relation

WordNet aims at organising the lexicon by semantic relations on the basis of synonymy. Synonymy is a semantic relation between two words with different forms and similar meanings. That is to say, the lexicon is organised in terms of word meaning rather than word forms. This mode of organisation makes WordNet thesaurus-like rather than dictionary-like. As a basic principle, meanings in WordNet are represented by synonym sets or synsets. A synset is the set of words that denote the same concept.

The relation between word meaning and word form in WordNet is characterised through a lexical matrix (see Figure 1). The matrix illustrates how word forms can be used to express word meanings, and a word form is polysemous or a synonym to another word form. F1 expresses word meaning M1. F1 and F2 are synonyms as they represent two entries in the same row. F2 is polysemous because it has two entries in the same column. The lexical matrix is based on the lexical semantic objective, which is mapping between forms and meanings i.e. it is represented through the actual mapping between written words and synsets.

Word meaning	Word form						
	F1	F2	F3	.	.	.	Fn
M1	E1.1	E1.2					
M2		E2.2					
M3			E3.3				
M4				.			
M5					.		
M6						.	
Mm							Em,

Figure 1: Lexical matrix

6 Sense relations

The thesaural relation of hyponymy is readily pictured in the relational model, as a transitive relation from synset to synset. Thus the hyponymy relation between the synsets entity,organism and organism, plant/flora is represented as in table 2. Separate tables store the instances of other sense relations in the same way, e.g. meronymy and antonymy. Hyponymy

Synset 1	Synset 2
100001742	105054818

Table 2: Representing hyponymy in a table

table and other similar tables showing transitive realtions are mainly used to support browsing related senses, as in Figure 2.

7 Adding data for Arabic and/or other languages

There are several alternative ways of adding a second and subsequent language to a sense enumerative lexicon, some, but not all of which are discussed in (Vossen et al., 1997). To make the

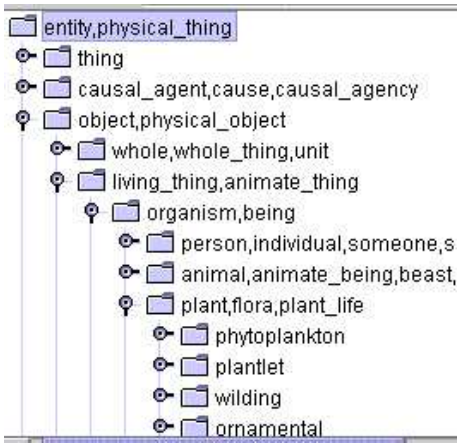

Figure 2: Tree viewer showing part of hyponymy relations

database multilingual, the basic need is to provide the equivalent of Table 1 for the additional language(s).

Three possible extensions to the data model suggest themselves:

(a)Change the name of the word column to English, and to add new columns for Arabic, French, etc.

Synset_No	W#	Eng	Arabic	Cat	S#
100001742	1	entity	wuju:d	n	i
100003135	1	organism	ka:in	n	1
103447508	1	plant	masna'	n	1
105054818	1	plant	naba:t	n	2
106962451	2	flora	naba:t	n	1
201241292	1	plant	zara'a	v	1

Table 3: adding a column to WN_S containing Arabic

(b)Add a new column in which a code for the language of the table row is placed.

(c)Reproduce WN_S table for each language.

An advantage of adding a new table is to make a new independent conceptual dictionary for the second language, whereas inserting a new column is more economical on space.

The Arabic equivalent of the WN_S table is created to include root and pattern of each word as additional columns as well as any language specific features. 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.

Synset_NO	W#	Word	Lang	Cat	S#
103447508	1	plant	English	n	1
103447508	1	masna'	Arabic	n	1
106962451	2	flora	English	n	1
106962451	2	naba:t	Arabic	n	1
105054818	1	plant	English	n	2
105054818	1	naba:t	Arabic	n	2
201241292	1	plant	English	v	1
201241292	1	zara'a	Arabic	v	1

Table 4: adding a column to WN_S containing language identification

Synset_No	Word	Cat	S#	Root	Pattern
103447508	masna'	n	1	s n '	maf'al
105054818	naba:t	n	2	n b t	fa'a:l
106962451	naba:t	n	1	n b t	fa'a:l
201241292	zara'a	v	1	z r '	fa'ala

Table 5: Arabic WN_S table

8 Querying translations

In either of the above database schemata, a translation query is straightforward, in one case requiring a join of two tables, in the other a single table query. We have joined WN_S table with WN_S_Arabic to show the English word, Arabic translation and the part of speech columns. For further clarity of senses we joined WN_G table to add the glosses and examples column to the query. The user or the intended user who is said to be the translator or the language specialist may prefer to leave the glosses in one language that can explain the sense of the word for both languages.

9 Updating translations

In an invirnment of an open-ended system for lexicon and terminology development, it will be critical to provide good facilities for entering translations and concepts and conceptual relations motivated by the second or subsequent language. In cases where Arabic words have no English translations, the following suggestions can be applied:

1- Allocate new Synset number. 2- Link Synset number to the nearest hypernym by adding row in WN_HYP table. 3- Add row

Synset_No	Word	Arabic	cat	Gloss
102837386	house	manzil	n	a dwelling that serves as living quarters
102838086	house	marab	n	a building in which something is sheltered
103491295	house	masrah	n	a building where theatrical performances can be presented
105976484	house	'aila	n	aristocratic family line.

Table 6: A join query of WN_S , WN_S_Arabic and WN_G tables

in WN_S table. 4- Add English gloss in the WN_GLOSS table.

10 Arabic Morphology

Arabic is highly inflectional language and can expand its vocabulary using a framework that is latent in the creative use of roots and patterns. Phonemes and letters are the components of the Arabic word. These components are mapped into a predetermined form known as the 'pattern'(Holes, 1995) to generate words. For example, the trilateral unagumented verbal root 'k t b 'can result in the following deivatives if subjected to certain patterns:

Arabic	English	POS	Pattern
kataba	write	v	fa'ala
kita:b	book	n	fi'a:l
kita:bah	writing	n	fi'a:lah
ka:tib	writer	n	fa:'il
ka:tib	clerk	n	fa:'il
ka:taba	correspond	v	fa:'ala
maktab	office	n	maf'al
maktabah	library	n	maf'alah
muka:tabah	correspondence	n	mufa:'alah
iktita:b	subscription	n	ifti'a:l
kita:bi	clerical	adj	fi'a:li

Table 7: deravatives of the arabic triliteral root k t b

It is worth mentioning that tables 7, 8 and 9 are for illustration purposes only and do not form a part of the database.

Consonants remain unchangeable and are not subjected to any conversion when deriving a new word, but they are derived from and built upon.

Grouping the sets of Arabic words according to their patterns will classify the language into distinct domains of nouns, verbs, adjectives and adverbs (Elkatib, 1991). This feature of Arabic is used in our design to query words from a give pattern to retrieve all Arabic words, their English translations, glosses examples and other related senses Table 8 shows different nouns coined according to the Arabic pattern taf'i:l which refers to a process or a progress of some activity:

Arabic word	English word
tasi:s	origination
tanzi:m	organization
ta'li:m	education
tajmi:'	assembly
takri:r	refining
tashhi:m	lubrication

Table 8: deravatives of the arabic triliteral root k t b

This feature of Arabic is also used to query Arabic words that are formed according to a given pattern to enable the language specialists to coin new Arabic terms accordingly.

Native Arabic speakers can easily tell the pattern of almost any given word, but also recall the words coined according to that pattern. In the data we have collected, there are lists of words that are searched according to given patterns. Every noun pattern for example is related to a particular verb. Therefore, in front of every noun in a list of a particular pattern there is a corresponding verb derived from the same root of that noun. It is important to note that those verbs listed are also coined according to a particular pattern. For example, the noun pattern 'tafa:'ul' Table 9 has a corresponding verb pattern 'tafa:'ala ':

11 User interface and the editing functionality

In order for the interface to satisfy all users who are or are not expected to have Arabic enabled version of Windows already installed, provide the functionality of Arabic script input mode in Java to support those with no AEW installed

Arabic noun	Meaning	Arabic verb
taba:dul	exchange	taba:dala
taba:'ud	separation	taba:'ada
tata:bu'	succession	tata:ba'a
taja:dhub	attraction	taja:dhaba
taqa:rub	approach	taqa:raba
taka:thur	multiplication	taka:thara
tama:thul	similarity	tama:thala
tana:fur	alienation	tana:fara
tana:fus	competition	tana:fasa

Table 9: finding noun verb relation through a given root

in their systems as well as for non-native speakers of Arabic who are welling to use systems and keyboards of their own languages. For this purpose a virtual keyboard is created, shown in Figure 3.

The interface uses information displays that treat each element as a distinct object rather than a text portion. All updates are made relative to an item previously retrieved, so the interface has a query facility. This allows words to be entered in either English or Arabic (and additionally Arabic roots and patterns), and a number of alternative queries invoked. Since words typically have multiple senses, the initial response to a query is to display a word sense matrix, shown in Figure 4.

Word_1	Word_2	Word_3
plant	flora	plant_life
plant	works	industrial_plant
plant		
plant		

Figure 4: Word Sense Matrix

The matrix allows cells, rows or columns to be selected. Selecting a cell or a row makes a particular synset current. This in turn enables the tree-view of a hierarchy of words to be generated and focused around the selected sense. At the same time, the gloss and examples for the selected sense are also retrieved and displayed.When a sense is selected either from the word sense matrix or from the tree viewer Arabic translation of the sense as well as root and pattern of the Arabic word are retrieved. Any updates are made relative to the synset currently shown as selected. See Figure 5.

12 Conclusion

The design and implementaion of the English-Arabic bilingual lexical resource is supported by a software framework together with a relational database populated initially with the contents of the WordNet.The design enables us to store more language specific lexicaland conceptual relations than those in the original wordnet. We will add further virtual relations, which can allow the conceptual dictionary to be augmented with morphological analysis and generation.

References

R. Beckwith and G.A. Miller. 1990. Implementing a lexical network. *International Journal of Lexicography 3*, pages 302–312.

S. M. Denness. 1996. A design of a structure for a multilingual conceptualdictionary. Msc dissertation, UMIST, Manchester, UK.

S. Elkatib. 1991. Translating scientific and technical information from english into arabic. Master's thesis, University of Salford, Manchester, UK.

C. Holes. 1995. *Modern Arabic.* Longman, London, UK.

G. A. Miller. 1990. Nouns in wordnet: Alexical inheritance system. *International Journal of Lexicography 3*, 4.

G. Sierra and J. McNaught. 2000. Design of an onomasiological search system: A concept-oriented tool for terminology. *Terminology*, 6(1):1–34.

P. Vossen, P. D?ez-Orzas, and W. Peters. 1997. The multilingual design of eurowordnet. In P. Vossen, N. Calzolari, G.Adriaens, A. Sanfilippo, and Y. Wilks, editors, *Proceedings of the ACL/EACL-97 workshop Automatic Information Extraction and Building of Lexical Semantic Resources for NLP Applications, Madrid, July 12th, 1997.*

Figure 3: Arabic Virtual Keyboard

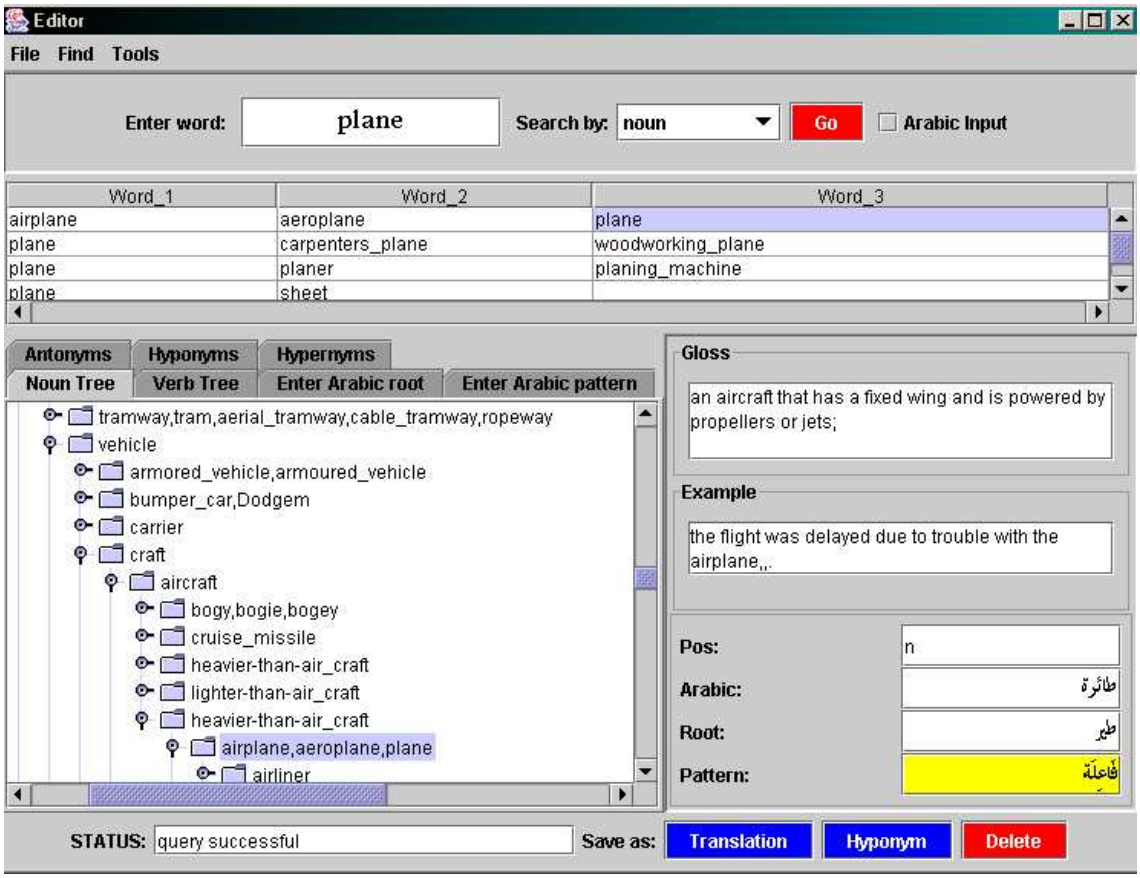

Figure 5: User's interface