The Need for a Reliable Dictionary

Dr. Sabri Elkateb
English Department - Faculty of Education
Zawia University

Abstract:

A dictionary is an aggregation of lexical items arranged in a certain order or a structure, with necessary information concerning them. This information may be expressed in a defining way within the same language to form a monolingual dictionary or a thesaurus, or in another language to form a dictionary for translation which can be either a bilingual or a multilingual dictionary. Any of these types of dictionaries cater for specific or several needs and may be presented in the form of a book, a card or may be computerized. This paper is an attempt to attract and to encourage further research in monolingual, bilingual and multilingual tools and resources for the Arabic language. The main focus here would be on exploiting more the features of Arabic and to devote more attention and research to the Arabic WordNet (AWN) discussed below.

1. Introduction:

Computers are now considered to be the only practical and efficient means of processing lexical and terminological data. Their storage capacity, speed and flexibility have attracted terminologists towards the automation of terminology and consequently lead to the creation of terminology data banks. Since there was no special processing system for lexical and terminological data, standard data processing methodologies were used. But due to the difference between the storage of lexical data and traditional information processing techniques, any results would be far from satisfactory. "Much progress is being made but further research is required in both conceptual modelling and computational lexicography/linguistics before anything approaching a complete lexical or terminological databank can be constructed". (Sager, 1996:129)

Computerized data dictionaries have become very valuable tools for the management of information resources. There are three types of data dictionaries, namely electronic word dictionaries which consist of only words, concept dictionaries that are characterized by a classification of hierarchy of relations and thesauruses in the form of semantic networks. A dictionary also has a thesaurus-based organization if semantic relations such as, synonymy, antonymy, hierarchical relations, etc. are included. This paper is an attempt to explore the field of computerized dictionaries with special focus on the computerized language tools that are considered as resources of lexical semantics. The main focus here would be on exploiting the features of Arabic and to devote more research to the Arabic monolingual and bilingual computerised dictionaries. Another aim is to attract and to encourage further research in monolingual, bilingual and multilingual tools and resources for the Arabic language.

2. Towards the Need for a Reliable Dictionary:

In the late 1960s, attempts were started to compute dictionaries for the purpose of language and computational explorations, but the resources available at that time could not cope with the huge amounts of language material to be explored. An example for that stage would be the *Webster's New Collegiate Dictionary*. However, the rapid advances in computer technology have encouraged more ambitions to come up with more valuable outcome. The well-known Collins Birmingham University International Language Database COBUILD project started in 1980 had succeeded in forming a computer database, which consists of a text corpus of about 20 million words in daily use. "What is new about this project, apart from the technology, is the ability to get for the first time a view of a language which is both broad and comprehensive. Many thousands of the observations are about the commonest patterns in the language". (Sinclair, 1987: vii)

The electronic versions of printed dictionaries, known as machine-readable dictionaries (MRDs), have encouraged researchers to exploit the ready-made data. In spite of the technical problems related to the fact that these MRD versions are written, in a specific way using codes and symbols, for human understanding not for a machine, large projects were based on paper dictionaries, like the Longman's Dictionary of Contemporary English (LDOCE). Despite some pessimistic views, researchers carrying out such projects view MRDs as sources of information useful for natural language processing (NLP). Their enthusiasm is triggered by their belief that information is relatively easy to extract from these MRDs. e.g., morphosemantic properties and lexical semantic information. Ide and Veronis (1993), representing a rather

pessimistic view, claim that the outcome is not worth all the efforts made, the result being a handful of limited and imperfect taxonomies. It seems that their views have encouraged researchers rather than impeding their enthusiasm. A fully automated dictionary in the form of a very large lexical knowledge base is being developed at Microsoft known as the "MindNet" using two MRDs: the Longman Dictionary of Contemporary English (LDOCE) and the American Heritage (AHD3, Third Edition).

Roget's International Thesaurus (RIT), introduces the English language as classification of words and has been described as a synonym dictionary. RIT was converted to MRD form in the early 1970's, and converted to a database form in the early 1990's. Its conceptual hierarchy and connectivity patterns compare to George Miller's WordNet, a model of the "mental lexicon" and its conceptual relations (Miller, et al 1993). WordNet is an alternative method of structuring *synonym sets*, based on psycholinguistic theories of human lexical memory, which has been attracting and encouraging research in the various fields of conceptual modelling and computational lexicography/linguistics, and in different languages of the world. Digital lexical resources can store lexicons of potentially unlimited size in ways that enable flexible representation and searches. Mapping the lexical inventory of a language into a semantic network has proven to be useful for many language processing applications. (Fellbaum & Vossen (2012)

3. WordNet, a Conceptual Lexical Resource:

Concepts are the organizational units in the WordNet model and they are more than a single word as they include compounds, collocations, idiomatic phrases, and phrasal verbs. "Compounds, collocations, idiomatic phrases, and phrasal verbs extend the idea of storing words in the lexicon to

storing conceptual information that may not have a lexical representation using a single word" (Jansen, 2004). One thing that WordNet does not do is to provide a topical organization of the lexicon (Miller, 1999).

WordNet is a widely available English lexical database and a valuable referable tool for language engineering and computational linguistics research. WordNet opened new insights for linguists and lexicographers and started a new era for computerised language tools and resources of lexical semantics. Its inspiration has given birth to several enormous monolingual and multilingual projects like the EuroWordNet and Balkanet covering fourteen European languages which, when linked through the Inter-Lingual-Index (ILI), will result in a huge network of linguistic concepts that allow inter-lingual navigation and search of translation equivalences between languages. WordNet-style lexicography has been applied to build resources in many languages. The Challenge we face now is to interconnect them so as to create one multilingual database. (Fellbaum & Vossen (2012)

WordNet as a lexical database offers broad coverage of the general lexicon in English. WordNet has been employed as a resource for many applications in information retrieval. Knowledge of words lies not only in their meanings but also in the context in which they occur. Linking words to appropriate senses provides the desired conceptual information. Terms holding identical meanings are organized around the notion of synonym set known as (synset) which represents a list of similar words that explain a concept. Synsets are linked to each other via pre-defined lexical relations. Furthermore, WordNet's high level classes have put some limit to enumeration of word senses keeping limited the search space of any generalization process.

4. Word-Sense Relation in WordNet:

Synonymy is a semantic relation between two words with different forms and similar meanings. The traditional way to define synonymy is in terms of substitution: Two words are synonyms (relative to a context) if there is a statement (or class of statements) in which they can be interchanged without affecting truth value (Miller, 1999).

In addition to synonymy, several other well defined semantic relations are recorded in the WordNet, namely, 1- hypernymy – hyponymy: a relation between the super-ordinate terms (hypernyms) and the subordinate terms (hyponyms), e.g. Parrot is a hyponym of a bird. 2- meronymy – holonymy: the part-whole semantic relation. A meronym is a word that names a part that belongs to a larger or more generic entity, e.g. a lens is a part of camera, while a holonym denotes a whole-part relation. 3- antonymy: word pairs that are opposite in meaning, such as fast and slow, rich and poor, 4- entailment : a relation occurs between two verbs having different senses, but logically one entails the other, like dance and move. 5- troponymy: the manner of action. To jog is to run in a certain manner. Then jog is a troponym of run..

Nouns participate in the relation of synonymy, antonymy, meronymy-holonymy, hypernymy-hyponymy relations, while verbs may be related by the synonymy, antonymy, troponymy, entailment, and hypernymy-hyponymy relations. Adjectives and adverbs are related by the synonymy and antonymy relations.

5. The Need for Bilingual Dictionaries:

Arabic is the official language of hundreds of millions of people in twenty Middle East and northern African countries, and is the religious language of all Muslims of various ethnicities around the world. Surprisingly, there is still little that has been done in the field of computerised language and lexical resources compared to large scale conceptual resources like the WordNet Black, et al (2006). It is therefore motivating to develop a WordNet-like lexical resource that discovers the richness of Arabic. An Arabic WordNet project based on Elkateb (2005) PhD Thesis is introduced to show the capability of Arabic to stand alone or be incorporated in a bilingual or a multilingual resource by the size and design of the English WordNet.

The aim has primarily been to design of a browsable and editable bilingual English-Arabic dictionary for the sake of comparing the two languages in order to tackle any language specific issues and solve any translation equivalence problems. Achieving that, then the ambition was the development of Arabic WordNet to be linked to the Global WordNet. The following figure shows the graphical user interface for the bilingual dictionary and the editing facility proposed by Elkateb (2005).

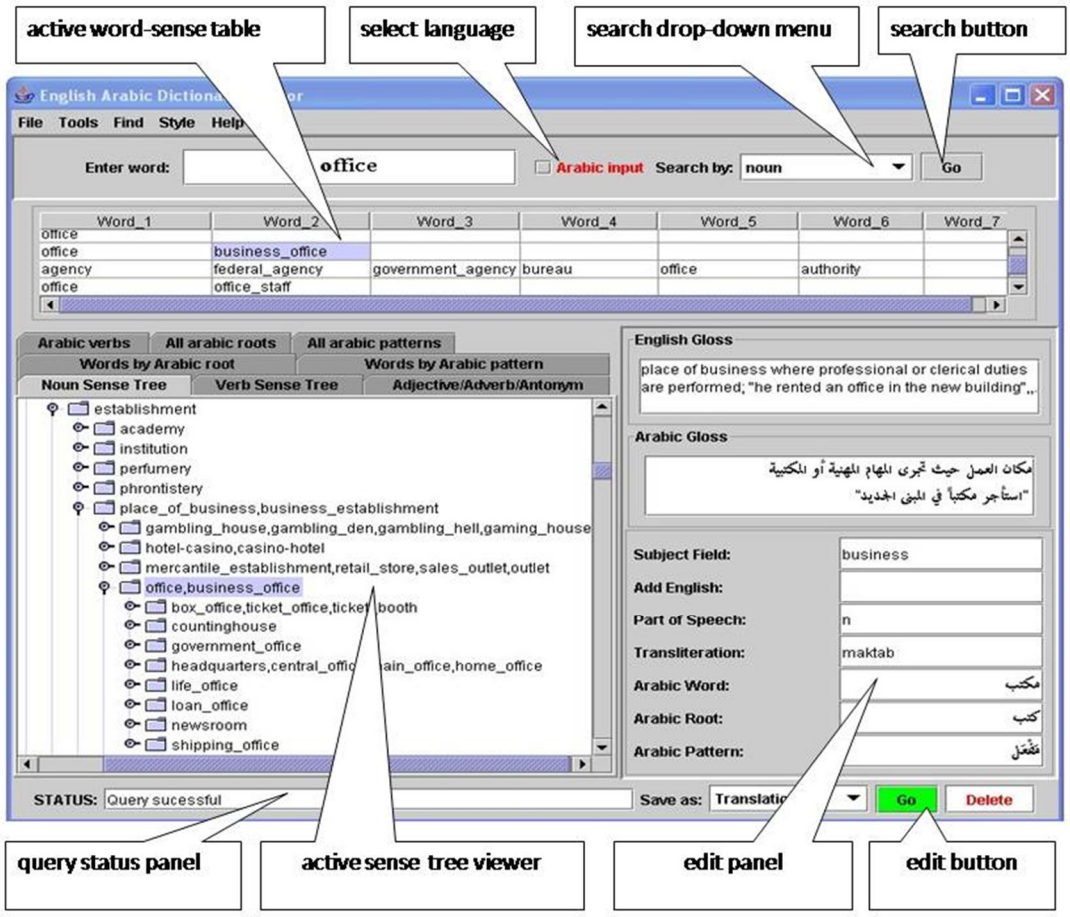

Figure (2) Editable dictionary interface showing Arabic features

As far as Arabic is concerned, the system provides several functionalities related to morphology which is designed to support queries that can find the following:

1. A root of a word.
2. Words derived from certain Arabic root.
3. A diacriticized pattern of the Arabic word.

4. Words that are coined according to a certain pattern.
5. Arabic verbs and their counterpart English equivalent verbs resulting from a derivation process.
6. All lexical relations proposed as in WordNet.

For a clearer picture of what both English WordNet and Elkateb's prototype Arabic dictionary/editor, the following figure shows results of both for the search word 'doctor'

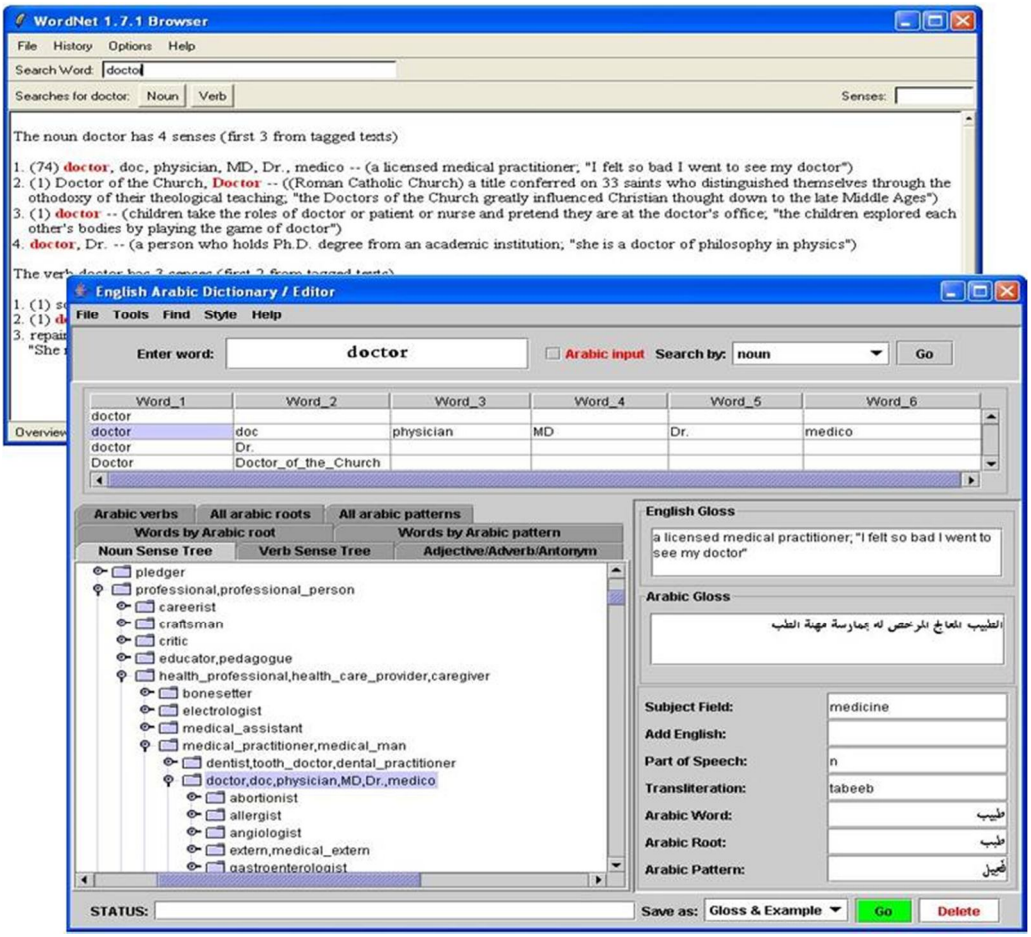

Figure (2) Comparing search results against WordNet

7. Utilization of Arabic Specific Features:

When considering languages more closely related to English, developing a multilingual wordnet can be as simple as providing the data mapping foreign words to the synsets of the Wordnet. Arabic, on the other hand, has an extensive system of derivational morphology that embodies important semantic relations, which ought to be reflected in any conceptual dictionary (Black & Elkateb, 2004).

An Arabic root is a kind of super-concept that is not a word but a skeleton from which to derive many words. For example, the word ‘kita:b’ (book) is derived from the necessary root ‘k t b’ with the addition of one short vowel ‘i’ and one long vowel ‘alif’. Most Arabic words are considered to be derived from roots formed by the insertion of short and long vowels and also some consonants like (m, t and n) added as affixes according to certain morphological patterns. The following table illustrates the functionality of Arabic roots, word forms and semantic relations generated by the Arabic root 'w l d' .

Arabic word		English word	Category
walada	وَلَدَ	give birth	v
wila:dah	وِلَادَة	birth	n
wallada	وَلَّدَ	generate	v
tawli:di	تَوْلِيدِي	generative	a
tawa:lada	تَوَالَدَ	breed	v
tawa:lud	تَوَالُد	reproduction	n
wa:lidah	وَالِدَة	female parent	n
wa:lid	وَالِد	male parent	n
walad	وَلَد	infant	n
wali:d	وَلِيد	born baby	n
wila:dah	وِلَادَة	delivery	n
wallada	وَلَّد	deliver/assist in birth	v
mawlid	مَوْلِد	Prophet’s birthday	n
mila:d	مِيلَاد	The Nativity	n
mi:la:di	مِيلَادِي	A.D., (anno domini)	n

Table (1) Word forms and semantic relations generated by the Arabic root 'w l d' .

Short vowels affect meaning, but are not normally written (but may be) (De Roeck and Al-Fares, 2000). Short vowels are conventionally

omitted in writing, but can be supplied by the reader, as they do not form part of the word and native Arabic writers and readers are familiar with the positions of such vowels, although absence of vowels in normal written texts makes some words apparently highly ambiguous out of context. “There is an assumption that roots and patterns define the meaning of lexical entries in Arabic.” (Cantineau, 1950; Cohen, 1961/70) cited in Dichy & Farghaly, (2003) Nouns, verbs and adjectives result from combination of (a) the “general meaning” of a given root, and (b) a “specific meaning” associated with a pattern.

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. In addition patterns can supply various additional features to a system. Those three components were incorporated in the design. However, it is not considered appropriate to attempt to `decode’ the patterns as semantic features or named relations.

8. Arabic WordNet (AWN):

AWN is constructed according to the methods developed for EuroWordNet (EWN); to maximize compatibility across wordnets and focuses on manual encoding of the most complicated and important concepts. Language-specific concepts and relations are encoded as needed or desired. This results in a so-called core wordnet for Arabic with the most important synsets, embedded in a solid semantic framework. From this core wordnet, it is possible to automatically extend the coverage with high precision. Specific concepts can be linked and translated with great accuracy because the base building blocks are manually defined and translated.

Constructing AWN presents challenges not encountered by established wordnets. These include the script on the one hand and the morphological properties of Semitic languages, centred around roots, on the other hand. The foundations for meeting these challenges have been laid. An innovation with significant consequences for wordnet development is the proposal to substitute English WordNet as the Interlingual Index (ILI) with the *Suggested Upper Merged Ontology* (*SUMO*) and its domain ontologies form the largest formal public ontology in existence today.

Accordingly with the objectives of the project, Arabic WordNet currently consists of 11,270 synsets (7,961 nominal, 2,538 verbal, 661 adjectival, and 110 adverbial), containing 23,496 Arabic expressions. This number includes 1,142 synsets that correspond to named entities which have been extracted automatically and are being checked by the lexicographers.

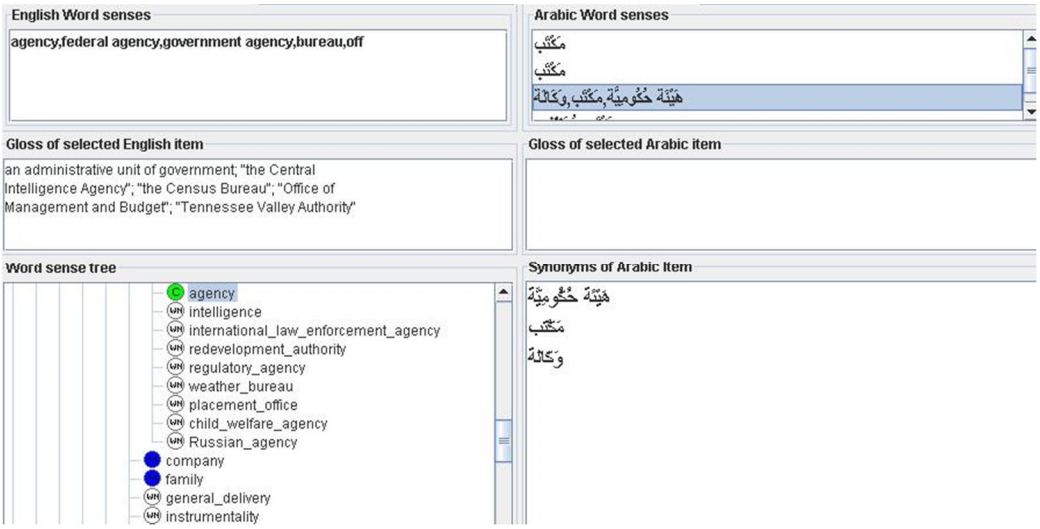

Figure (3) Arabic WordNet (AWN) Browser

A promising research focuses on the semi-automatic extension of Arabic WordNet (AWN) using lexical and morphological rules. (Rodríguez, et al 2008) There has always been a need for extending AWN coverage by taking advantage of a limited set of highly productive Arabic morphological rules for deriving a range of semantically related word forms from verb entries. Recent research on AWN is carried out by Abdurrahim, et al. (2013) implementing a method of semantic indexing of documents and query for the information retrieval where AWN is used as a semantic resource to explore the impact of a passage from an indexation based on single words to an indexation based on concepts.

9. Conclusion:

Any lexical or terminological database is expected to accept a variety of queries. There are queries about single items of data like grammatical category, synonym, antonym, hyponym, definition of a term, etc., or about a set of data like a list of terms that are nouns of X, or lists of terms in a subject area Y, etc. Searches or queries can also be made about a term or terms that are not known, where information about their meaning is known, i.e. we can arrive at a term or a list of terms by means of their synonyms or definitions. Wordnets provide the sources of data and a clear framework for lexical representation in addition to their importance as resources for many applications within language technology. They can be used in meaning-based information retrieval (searching for concepts rather than specific word forms), in logical inference (if a document mentions dogs, a wordnet allows the inference that it is about animals), in word sense disambiguation (providing the search space of alternative meanings), etc. (Dyvik, 2002) Wordnets have been created in many languages, revealing both their lexical

commonalities and diversity. The next challenge is to make multilingual wordnets fully interoperable. (Fellbaum & Vossen (2012)

References:

1. *Abderrahim, M. A., Abderrahim, E. and Chikh, M. A. (2013)* ***Using Arabic Wordnet for semantic indexation in information retrieval system.*** *IJCSI International Journal of Computer Science Issues, Vol. 10, Issue 1, No 2, January 2013.*

2. *Al- Kasimi, A. (1983)* ***Linguistics and Bilingual Dictionaries,*** *Leiden E. J. Brill.*

3. *Black, W. J. & Elkateb, S. (2004)* ***A Prototype English-Arabic Dictionary Based on WordNet****, Proceedings of 2nd Global WordNet Conference, GWC2004, Czech Republic: 67-74*

4. *Black, W., Elkateb, S., Rodriguez, H, Alkhalifa, M.,Vossen, P., Pease, A., and Fellbaum, C.(2006).* ***Introducing the Arabic WordNet Project****. In Proceedings of the Third International WordNet Conference*

5. *De Roeck, A. and Al-Fares, W. (2000)* ***A Morphologically Sensitive Clustering Algorithm for Identifying Arabic Roots****, Proceedings of the 38th Annual Meeting of the ACL, Hong Kong:199-206*

6. *Dichy, J. and Farghaly, A. (2003)* ***Roots & Patterns vs. Stems plus Grammar-Lexis Specifications****: Paper presented at the: Ixth MT*

Summit – Workshop on Machine Translation for Semitic Languages: issues and approaches – New Orleans, USA

7. *Elkateb, S and Black, W (2001)* ***Towards the Design of English-Arabic Terminological Knowledge Base****. Proceedings of ACL 2000, Toulouse, France:113-118*

8. *Elkateb, S. (2005)* ***Design and Implementation of an English-Arabic Dictionary Editor****, PhD Thesis submitted to the University of Manchester 2005.*

9. *Elkateb, S., Black, W., Rodríguez, H., Alkhalifa, M., Vossen, P., Pease, A. and Fellbaum, C.: (2006)* ***Building a WordNet for Arabic****, in Proceedings of The fifth international conference on Language Resources and Evaluation; Genoa-Italy, 2006, pp 29-34.*

10. *Elkateb, S., Black, W., Vossen, P., Pease, A. and Fellbaum, C. (2006) Arabic WordNet and the Challenges of Arabic for NLP/MT. International conference at the British Computer Society, London, 2006; pp 15-24.*

11. *Elkateb. S. & Black, W (2004)* ***A Bilingual Dictionary with Enriched Lexical Information****, Proceedings of NEMLAR Cairo, Egypt 2004 Arabic Language Tools and Resources:79-84*

12. *Fellbaum, C. & Vossen, P. (2012)* ***Challenges for a multilingual wordnet****. Language Resources and Evaluation Vol. 46 No. 2 Pg. 313-32*

13. *Ide, N. and J. Veronis. (1993)* ***Extracting knowledge bases from machine-readable dictionaries****: have we wasted our time? KB & KS, December, 1993, Tokyo*

14. *Jansen, P. (2004)* ***Lexicography in an Interlingual Ontology****: An Introduction to EuroWordNet, Canadian Undergraduate Journal of Cognitive Science, 2004 vol ii:1-5*

15. *Matthews, P. H. (1974)* ***Morphology****: An Introduction to the Theory of Word Structure. Cambridge University Press*

16. *Miller, G. A. (1999)* ***ON KNOWING A WORD****, Annual Review of Psychology,* ***50****, 1-19.*

17. *Miller, G., Beckwith, R., Fellbaum, C., Gross, D., Miller, K. & Tengi, R. (1993) Introduction to WordNet:* ***An On-line Lexical Database****, WordNet Five papers on WordNet, Princeton University*

18. *Piasecki, M., Szpakowicz, P., Fellbaum, C. and Pedersen, B. (2013)* ***Introduction to the Special Issue: On Wordnets and Relations.*** *Language Resources and Evaluation Vol. 47 No. 3 Pg. 757-767*

19. *Rodríguez, R., Farwell, D., Farreres, J., Bertran, M., Alkhalifa,M., Martí, M.A., Black,W., Elkateb, S., Kirk, J., Pease, A., Vossen, P., and Fellbaum, C. (2008).* ***Arabic WordNet: Current State and Future Extensions****. Proceedings of The Fourth Global WordNet Conference, Szeged, Hungary. January 22-25, 2008.*

20. *Sager, J. (1996)* ***A Practical Course In Terminology Pro****cessing. John Benjamins B. V.*

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

22. *Sinclair, J. M. (1987)* ***Looking Up****. An Account of the Cobuild Project in Lexical Computing, Collins*