Arabic WordNet and the Challenges of Arabic

Sabri Elkateb, William Black
The University of Manchester
PO Box 88, Sackville St, Manchester, M60 1QD
sabri.elkateb@manchester.ac.uk
w.black@manchester.ac.uk,
Piek Vossen
Irion Technologies
Irion Technologies, Delftechpark 26, 2628XH, Delft, The Netherlands
piek.vossen@irion.nl

David Farwell
Politechnical University of Catalonia
Jordi Girona, 1-3, 08034 Barcelona, SPAIN
farwell@lsi.upc.edu
Adam Pease
Articulate Software Inc, 278 Monroe Dr. #30
Mountain View, CA 94040
apease@articulatesoftware.com
Christiane Fellbaum
Princeton University, Department of Psychology,
Green Hall, Princeton, NJ 08544
fellbaum@clarity.princeton.edu

Arabic WordNet is a lexical resource for Modern Standard Arabic based on the widely used Princeton WordNet for English (Fellbaum, 1998). Arabic WordNet (AWN) is based on the design and contents of the universally accepted Princeton WordNet (PWN) and will be mappable straightforwardly onto PWN 2.0 and EuroWordNet (EWN), enabling translation on the lexical level to English and dozens of other languages. We have developed and linked the AWN with the Suggested Upper Merged Ontology (SUMO), where concepts are defined with machine interpretable semantics in first order logic (Niles and Pease, 2001). We have greatly extended the ontology and its set of mappings to provide formal terms and definitions for each synset. The end product would be a linguistic resource with a deep formal semantic foundation that is able to capture the richness of Arabic as described in Elkateb (2005). Tools we have developed as part of this effort include a lexicographer's interface modeled on that used for EuroWordNet, with added facilities for Arabic script, following Black and Elkateb's earlier work (2004). In this paper we describe our methodology for building a lexical resource in Arabic and the challenge of Arabic for lexical resources.

1. Introduction

Arabic WordNet is being constructed following methods developed for EuroWordNet (Vossen, 1998). EuroWordNet approach maximizes 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. The approach follows a top-down procedure. Arabic Base Concepts are defined and extended via hyponymic relations to derive a core wordnet. The set of Common Base Concepts (CBCs) from the 12 languages in EWN and BalkaNet (Tufis 2004) are encoded as synsets; other language-specific concepts are added and translated manually to the closest synset(s) in Arabic. The same step is performed for all English synsets that currently have an equivalence relation in SUMO.

The first layers of hyponyms are chosen on the basis of linguistic and applications based criteria; the final phase completes the target set of concepts/synsets, including specific domains and named entities. Each synset construction step is followed by a validation phase, where formal consistency is checked and the coverage is evaluated in terms of frequency of occurrence and domain distribution.

2. WordNet

WordNet as a lexical resource offers broad coverage of the general lexicon. 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 a synset. 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. Concepts are the organizational units in the WordNet 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).

The success of the Princeton WordNet (PWN) for English has motivated similar projects that aim at developing wordnets for other languages. In recent years, a number of wordnet building efforts have been initiated and carried out within a common framework for lexical representation and are becoming increasingly important resources for a wide range of Natural Language Processing applications. Another promising initiative is the foundation of the Global WordNet Association when it was discovered that "WordNet had caught on around the world. Not only were foreign linguists building their own versions, but commercial companies were using the dictionary for their own purposes because it was available free via the World Wide Web". (Heyboer, 2002) The Global WordNet project aims to coordinate the production and linking of wordnets for all languages of the world.

3. Constructing AWN

The basic criteria for selecting synsets covered in AWN are:

- **Connectivity:** AWN should be as densely connected as possible by hyperonymy/ hyponymy chains, etc. Most of the synsets of AWN should correspond to English WN counterparts and the overall topology of both wordnets should be similar.
- **Relevance:** Frequent and salient concepts have priority. Criteria will include the frequency of lexical items (both in Arabic and English) and the frequency of Arabic roots in their respective reference corpora.
- **Generality:** Synsets on the highest levels of WN are preferred.

These criteria suggest two ways for proceeding:

- **From English to Arabic:** Given an English synset, all corresponding Arabic variants (if any) will be selected.
- **From Arabic to English:** Given an Arabic word, all its senses have to be found, and for each of these senses the corresponding English synsets have to be selected.

4. Challenges of Arabic

4.1. Target Users

Arabic is a Semitic language which differs from Indo-European languages syntactically, morphologically and semantically. The term 'classical Arabic' refers to the standard form of the language used in all writing and heard on television, radio and in public speeches and religious sermons. The writing system of Arabic has twenty five consonants and three long vowels that are written from right to left and take different shapes according to their position in the word. In addition to the long vowels, Arabic has short vowels. Short vowels are not part of the alphabet but rather are written as vowel diacritics above or under a consonant to give it its desired sound and hence give a word a desired meaning. Texts without vowels are considered to be more appropriate by the Arabic-speaking community since this is the usual form of everyday written and printed materials (books, magazines, newspapers, letters, etc.). But when it comes to the text of the Holy Koran, and more generally to printed collections of classical poetry, school books and some Arabic paper dictionaries, vowel diacritics appear in full. It is very usual for well-edited books, some printed texts, and manuscripts to have vowel diacritics partially or randomly written out in cases where words will be ambiguous or difficult to read. For instance, a word in Arabic consisting of three letters like (علم) can be very ambiguous without vowel diacritics. Consider the examples in Table 1. Especially in such cases as these, a writer may use diacritics so readers can easily resolve any ambiguity. However, although most Arabs can read texts with vowels explicitly indicated, fewer can write texts using the correct vowel diacritics.

Arabic	Transliteration	PoS	meaning
عَلَم	'alam	n	flag
عِلْم	'ilm	n	science
عُلِمَ	ulima	v	known
عَلَّمَ	'allama	v	teach
عَلَم	'alam	a	famous

Table 1: Vowel diacritics

For this reason it is a mistake to rely on users, regardless of their background, to correctly enter a search word requiring vowel diacritics. Yet misuse of a single diacritic, such as the '*suku:n*' which indicates that a consonant is not followed by any vowel, or as the '*shaddah*' (as in *'allama* in Table 1 and *darrasa* in Table 2), which indicates a double consonant, will cause a query to fail. People also tend to make mistakes about the position of some diacritics in a word. This can pose a serious problem for information retrieval systems and computerized lexical resources which depend on well-formed user input and may even result in users rejecting the system. In particular, there may be an outright rejection of a robust new lexical resource such as AWN unless that new resource assumes that most of the Arabic speaking users do not have expert command in writing vowel diacritics and will generally ignore them. These users are more comfortable reading texts without diacritics in dealing with everyday written materials including legal and business contracts, newspapers, books as well as both paper and computerized dictionaries. The end result is that it is preferable to allow users

to enter Arabic words without diacritics while at the same time allowing the retrieval of those words with vowel diacritics for the purposes of disambiguation.

Another fact about Arabic to take into consideration is that the language has no capital letters (for proper names: the names of people, countries, cities, geographical features, of months, days of the week, etc.) makes scant use of acronyms. This creates increased ambiguity and especially complicates such tasks as Information Extraction in general and Named Entity Recognition in particular.

4.2. Arabic Morphology

An additional property of Arabic that should be kept in mind is that Arabic is a highly derivational and inflectional language and its vocabulary can be easily expanded using a framework that is latent in the creative use of roots and morphological patterns. According to Al-Fedaghi and Al-Anzi (1989), cited in De Roeck and Al-Fares (2000), "85% of words derived from tri-literal roots" and there are around 10.000 independent roots.
Because of this, it is possible to build any necessary semantic relation among words of different syntactic categories. That is to say, most Arabic words are created by applying distinct derivational patterns to some root, relating the two not only in form and meaning but determining their syntactic category as well. New Arabic words can always be coined from an existing root according to the standard derivational patterns. It is also possible to organize sets of Arabic words into distinct semantic fields according to the root from which they are derived.

4.3. Processing of Arabic morphology

Numerous efforts have been devoted to the processing of Arabic morphology which outcome is apparent in several approaches and various technical morphological analysers and generators. Among other computational approaches to Arabic morphology, using techniques of Finite State Transducer (FST) and two-level morphology is Beesley (1998, 2001) His system dealt with root, stem and pattern morphology using only two layers. One layer corresponds to the root and is represented by the root lexicon and the other to the morphological measure including vowel pattern.

However, in order to produce a system on the basis of morphological analysis and generation that is linguistically and computationally efficient; the following factors have to be taken into consideration:

1. A word pattern usually combines with a vast number of roots. Roots and patterns are intersected at compile time to yield 90,000 stems. Various combination of prefixes and suffixes, concatenated to the stems, yield over 72,000,000 abstract words.
2. The existence of one morphological form depends on the existence of other forms comprised of the same morphological unit.
3. There are cases where a single form has more than one morphological function as illustrated in Table 1 above.
4. A word is generated by the combination of a root encoded manually and a diacritized pattern each of which has to be hand coded to indicate the subset of patterns with which a root can combine.
5. A root can be extracted by removing the affixes to identify the base form of the diacritized word and to apply it to a morphological measure or a pattern. In this case both word and pattern must be entered manually.

6. Some techniques are designed not to take any Arabic text as an input directly, but to transliterate the Arabic system into ASCII to be fed to the system. The results must be transliterated back to Arabic to be understood. This technique was introduced by Buckwalter (2002) and can be said to have achieved considerable results in Arabic morphological analysis, yet it is unable to adequately deal with ambiguous forms but can only provide full listing of all the possible readings of the ambiguous form.

There seems to be no agreement on the nearest way to adequate morphological analysis/generation and there is yet no proper means for generating or analyzing the Arabic roots due to the complexity of the weak vowels governing a vast amount of the vocabulary. It seems also that there is no role for morphological generation in suggesting words, because for much of the vocabulary, the rate at which these would prove to be actual words would be too low unless at least three quarters of the process are done manually (Elkateb, 2005).

4.4. Language specific and untranslatable material

As far as dictionaries are concerned, a multilingual resource generally includes equivalence and translation relations and should tackle issues like language specific and untranslatable material. Translation is not merely an act of linguistic transfer, but it also involves the interaction of cultures and that transference of culture imposes far greater problems than linguistic transfer. Translation of words of cultural content may involve solving problems like the unavailability of equivalents or tackling untranslatable items and consequently filling the gaps that may exist among languages. Consider the examples in Table 2:

عيد الفطر	*?eid alfitr*	The socio religious event in which Muslims celebrate their end of fasting in the Holy month of Ramadan.
أضحية	*udHiya*	a sheep killed as sacrifice on the day of The Greater ?eid.
سحور	*Suhu:r*	a light meal before starting a new day of Ramadan (before daybreak).
زكاة	*Zaka:t*	an annual compulsory alms (2.5 %) of the savings of a Muslim when any amount or property exceeds

Table 2: Lexical gaps

- **Synonymy and confusion of non-standardised terms (for the translator)**

Thermometer: Arabic equivalents in a paper dictionary

- miHarr مِحَرّ
- miHra:r مِحْرَار
- miqya:s Hara:rah مقياس حرارة
- miza:n Hara:rah ميزان حرارة

- termomitr ترمومتر

- **Technical translation: precision and economy (Elkateb 1991)**
  - ***reboiler:*** Galla:yat I'a:dat alGali wattabkhi:r غلاية إعادة الغلي والتبخير (cannot be reduced)
  - **hydrometer:** is rendered in Arabic by long phrases **,** jiha:z qiya:s kathafat assawa:il جهاز قياس كثافة السوائل (cannot be reduced
  - **Hydrogenation:** 'mu'aalajah bilhaydroji:n' معالجة بالهيدروجين is reduced (standardized) to *'hadrajah'* هدرجة as in 'aksadah' أكسدة for oxidization.

- English word *uncle* (*the brother of your father or mother; the husband of your aunt)*

  Arabic uses a different word for each member as a separate concept like '*amm'* عَمّ ' for the father's brother and خَال '*khal*' for the brother of the mother, zawj al'ammah:زوج العمة the husband of your aunt (the sister of your father) and zawj alkhalah:زوج الخالة the husband of your aunt (the sister of your mother).

- English word *parent*: a father or mother; one who begets or one who gives birth to or nurtures and raises a child; a relative who plays the role of guardian

  Arabic uses *walid* والدfor male parent and *walidah* والدة for a female parent (use of relative does not exist)

- English word *spouse , partner , married person , mate , better half* (a person's partner in marriage)

  Arabic uses one form for masculine and same form + ta marbuta ـة for feminine

masculine	feminine
zawjزوج	zawjat زوجة
Shari:k Haya:tشريك حياة	Shari:kat Haya:t شريكة حياة
Qari:n قرين	Qari:nat قرينة
Hali:lحليل	Hali:lat حليلة
Ba'lبَعْل	Ba'lat بَعْلة
	'aqi:lat عقيلة

Table 3: Use of feminine in Arabic for spouse

5. Lexicography

Following EuroWordNet, AWN is developed in two phases by first building a core wordnet around the most important concepts, the so-called Base Concepts (Vossen 1998), and secondly extending the core wordnet downward to more specific concepts using additional criteria. The core wordnet should thus become highly compatible with wordnets in other languages that are developed according to the same approach.

For the core wordnet, The Common Base Concepts (CBCs) of the 12 languages in EWN and BalkaNet (Tufis, 2004) are being encoded as synsets in AWN; other Arabic language-specific concepts are added and translated manually to the closest synset. The same procedure is performed for all English synsets that currently have an equivalence

relation in the SUMO ontology. Synset encoding proceeds bi-directionally: given an English synset, all corresponding Arabic variants (if any) will be selected; given an Arabic word, all its senses are determined and for each of them the corresponding English synset is encoded.

The Arabic synsets will be extended with hypernym relations to form a closed semantic hierarchy. SUMO is used to maximize the semantic consistency of the hyponymy links. This will represent the core wordnet, which is a semantic basic for the further extension. The work is mostly done manually.

When a new Arabic verb is added, extensions are made from verbal entries, including verbal derivates, nominalizations, verbal nouns, and so on. We also consider the most productive forms of deriving broken plurals. This is done by applying lexical and morphological rules iteratively.

The database is further extended downward from the CBCs. First, a layer of hyponyms is chosen based on maximal connectivity, relevance, and generality. Two major pre-processing steps are required, preparation and extension. Preparation entails compiling lexical and morphological rules and processing available bilingual resources from which we construct a homogeneous bilingual dictionary containing information on the Arabic/English word pair. This information includes the Arabic root, the POS, the relative frequencies and the sources supporting the pairing. The Arabic words in these bilingual resources must also be normalized and lemmatized while maintaining vowels and diacritics.

We next apply 17 heuristic procedures, previously used for EWN, to the bilingual dictionary in order to derive candidate Arabic words/English synsets mappings. Each mapping includes the Arabic word and root, the English synset, the POS, the relative frequencies, a mapping score, the absolute depth in AWN, the number of gaps between the synset and the top of the AWN hierarchy, and attested tokens of the pair. The Arabic word/English synset pairs constitute the input to a manual validation process. We proceed by chunks of related units (sets of related WN synsets, e.g. hyponymy chains and sets of related Arabic words, i.e., words having the same root) instead of individual units (i.e., synsets, senses, words).

Finally, AWN will be completed by filling in the gaps in its structure, covering specific domains, adding terminology and named entities, etc. Each synset construction step is followed by a validation phase, where formal consistency is checked and the coverage is evaluated in terms of frequency of occurrence and domain distribution. The total coverage of AWN will be around 10,000 synsets.

6. Tools

A lexicographer's interface modeled on the EWN interface with added facilities for Arabic script is being developed. Because AWN is to be aligned not just to PWN but to every wordnet aligned to PWN –either directly or indirectly through an Interlingual Index or the ontology – the database design supports multiple languages. The user interface will be explicitly multilingual and indifferent to the direction of alignment between the conceptual structures of the two languages. In addition to search and browsing facilities for the end users of the completed database, lexicographers require an editing interface. A variety of legacy components are available, each with their

relative advantages. The editor's interface communicates with the database server using Simple Object Access Protocol (SOAP), allowing multiple lexicographers at different sites to maintain a common database.

7. Database Structure

The database structure comprises four principal entity types: *item*, *word*, *form* and *link*. *Items* are conceptual entities, including synsets, ontology classes and instances. An item has a unique identifier and descriptive information such as a gloss. Items lexicalized in different languages are distinct. A *word* entity is a word sense, where the word's citation form is associated with an item via its identifier. A *form* is an entity that contains lexical information (not merely inflectional variation). The forms are the root and/or the broken plural form, where applicable. A *link* relates two items, and has a *type* such as "equivalence," "subsuming," etc. Links interconnect sense items, e.g., a PWN synset to an AWN synset, a synset to a SUMO concept, etc. This data model has been specified in XML as an interchange format, but is also implemented in a MySQL database hosted by one of the partners.

8. Ontology

A large ontology providing the semantic underpinning for AWN concepts is built on SUMO, a formal ontology of about 1000 terms and 4000 definitional statements currently that is provided in a first order logic language called Standard Upper Ontology Knowledge Interchange format (SUO-KIF) and also translated into OWL semantic web language. SUMO has natural language generation templates and a multi-lingual lexicon that allows statements in SUO-KIF and SUMO to be expressed in multiple languages. Synsets map to a general SUMO term or a term that is directly equivalent to the given synset (Figure 1).

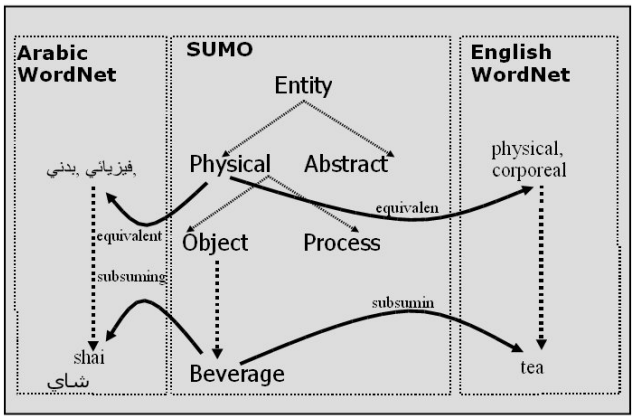

Figure 1: SUMO mapping to wordnets

New formal terms will be defined to cover a greater number of equivalence mappings, and the definitions of the new terms will in turn depend upon existing fundamental concepts in SUMO. The process of formalizing definitions will generate feedback as to whether word senses in AWN need to be divided or combined and how glosses may be clarified. Wordnets in other languages linked by synset number will benefit, too. The Sigma ontology development environment will be updated to handle a similar presentation of Unicode-based character sets, including Arabic.

The Interlingual Index (ILI) connecting EWN wordnets is a condensed set of more or less universal concepts linking synsets across languages via multiple exhaustive

equivalence relations. In EuroWordNet and BalkaNet, English PWN has been used to express equivalence relations across the different languages. By providing many SUMO definitions and terms that correspond to Arabic synsets, we will create the opportunity to use SUMO as the ILI for all wordnets that are currently related to PWN. This is illustrated in Figure 2. If the Arabic word sense for *shai* is exhaustively defined by relations to SUMO terms, this definition can replace an equivalence relation (er1) that is currently encoded between the Arabic synset *shai* and a synset *tea* in PWN. Note that the relations from *shai* to the SUMO terms need to be exhaustive, which may require multiple relations of different types (sr1 (subsumption), r2, r3) to multiple SUMO terms.

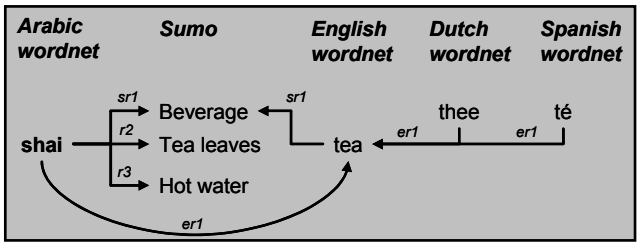

Figure 2: SUMO and ILI

If there are also equivalence relations from other languages (e.g. Dutch and Spanish) to the same PWN synset, then these relations grant the linkage of the synsets in these languages to the same SUMO definition.

Besides providing a formal semantic framework, SUMO can thus also be used to map synsets across languages, in fact even when there is not an equivalent in English. By composing formal definitions for the non-English synsets, SUMO as an ILI will not only be less biased by English but also has more expressive power.

9. Conclusion

Building an Arabic wordnet with the qualities discussed above presents challenges not encountered by established wordnets. These include the script on the one hand and the morphological properties of Semitic languages, centered 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 WN as the ILI with SUMO.

10. Acknowledgements

This work was supported by the United States Central Intelligence Agency.

11. References

Beesley, K. (2001) Finite-State Morphological Analysis and Generation of Arabic at Xerox, ACL/EACL 2001, July 6th, Toulouse, France : 1-8

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, Sojka, Choi, Fellbaum and Vossen eds.

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

Buckwalter, T. (2002) Arabic Morphological Analysis, Http://www.qamus.org/morphology.htm

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

Dyvik, H. (2003) Translations as a semantic knowledge source: word alignment and wordnet, Section for Linguistic Studies scientific papers, University of Bergen

Dyvik, H. (2002) Translations as Semantic Mirrors: From Parallel Corpus to Wordnet1. Section for Linguistic Studies scientific papers, University of Bergen

Elkateb, S. (2005) Design and implementation of an English Arabic dictionary/editor. PhD thesis, The University of Manchester, United Kingdom.

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

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

Elkateb, S. (1991) Translating Scientific and Technical Information from English into Arabic, Salford University.

Farreres, J. (2005) Creation of wide-coverage domain-independent ontologies. PhD thesis, Univertitat Politècnicade Catalunya.

Fellbaum, C., (1998, ed.) WordNet: An Electronic Lexical Database. Cambridge, MA: MIT Press.

Heyboer, B. (2002) Wizard of the New Wordsmiths, His idea to link words rewrote the dictionary, Interview with Miller, G. in Star-Ledger Newspaper, Tuesday, January 22, 2002 http://www.nj.com/news/ledger/index.ssf?/page1/ledger/15a30b1.html

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

Niles, I., and Pease, A. (2001) Towards a Standard Upper Ontology. In: Proceedings of FOIS 2001, Ogunquit, Maine, pp. 2-9.

Pease, A., (2000) Standard Upper Ontology Knowledge Interchange Format. Web document http://suo.ieee.org/suo-kif.html.

Pease, A., (2003) The Sigma Ontology Development Environment, in Working Notes of the IJCAI-2003 Workshop on Ontology and Distributed Systems. Volume 71 of CEUR Workshop Proceeding series

Pustejovsky, J. (1995) The Generative Lexicon, Massachusetts Institute of Technology.

Tufis, D. (ed.) (2004) Special Issue on the BalkaNet project. Romanian Journal of Information Science and Technology, Vol.7, nos 1-2

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

Vossen P. (2004) EuroWordNet: a multilingual database of autonomous and language-specific wordnets connected via an Inter-Lingual-Index. International Journal of Lexicography, Vol.17 No. 2, OUP, 161-173.