Procesamiento del Lenguaje Natural, Revista nº 44, marzo de 2010, pp 19-26 recibido 08-01-10 revisado 21-02-10 aceptado 01-03-10

Enhancing UniArab with FunGramKB

Cómo mejorar UniArab con FunGramKB

Carlos Periñán-Pascual
Universidad Católica San Antonio
Departamento de Idiomas
Campus de los Jerónimos s/n
30107 Guadalupe - Murcia (Spain)
jcperinan@pdi.ucam.edu

Ricardo Mairal Usón
Universidad Nacional de Educación a Distancia
Facultad de Filología
Senda del Rey, 7 - 28040 Madrid (Spain)
rmairal@flog.uned.es

Resumen: En el marco de la traducción automática árabe-inglés, el enfoque pseudointerlingüístico de UniArab ha logrado, incluso con oraciones simples, mejores resultados que los traductores automáticos basados en modelos estadísticos. El éxito de UniArab se cimienta en el modelo funcional de la Gramática del Papel y la Referencia, la cual es capaz de reconstruir la estructura lógica subyacente a un texto de entrada. No obstante, es preciso reemplazar la base de datos léxica de este traductor automático por una base de conocimiento más robusta con el fin de procesar textos lingüísticamente más complejos. De hecho, la integración de FunGramKB en la arquitectura de UniArab permite que este traductor automático utilice ahora una auténtica representación interlingüística denominada "estructura lógica conceptual", dando lugar a un enfoque conceptualista que favorece la generación multilingüe.

Palabras clave: traducción automática, UniArab, FunGramKB, base de conocimiento, estructura lógica, interlingua

Abstract: In the field of the Arabic-to-English machine translation, the pseudo-interlingual approach of UniArab clearly outperforms existing statistical machine translators, even only with the processing of simple sentences. The success of UniArab is founded upon the functional model of Role and Reference Grammar, which is able to reconstruct the logical structure underlying the input. However, it is essential to replace the UniArab lexical database with a robust knowledge base which enables linguistically-complex texts to be processed adequately. Indeed, the integration of FunGramKB into the architecture of UniArab allows the system to use a real interlingual representation known as "conceptual logical structure", resulting in a conceptualist approach which supports multilingual generation.

Keywords: machine translation, UniArab, FunGramKB, knowledge base, logical structure, interlingua

1 *Introduction*

The multilingual communities in Europe, together with the migration of people in today's world, has re-enforced the urgent need for a shared understanding of information. Therefore, the need for language-aware software applications that are grounded in a robust linguistic model is critical to our society. In this context, Arabic has become a major focus of interest. The aim of this paper is to describe how the machine translator UniArab can be dramatically enhanced by the integration of the knowledge base FunGramKB.

2 *FunGramKB*

FunGramKB[1] (Periñán-Pascual and Arcas-Túnez, 2004, 2005, 2007; Mairal Usón and Periñán-Pascual, 2009; Periñán-Pascual and Mairal Usón, 2009) is a multipurpose lexico-

[1] www.fungramkb.com

ISSN 1135-5948

conceptual knowledge base for natural language processing systems, and more particularly for natural language understanding. On the one hand, FunGramKB is multipurpose, in the sense that it is both multifunctional and multilingual. Thus, FunGramKB has been designed to be potentially reused in many natural language processing tasks (e.g. information retrieval and extraction, machine translation, dialogue-based systems, etc) and with many natural languages.[2] On the other hand, our knowledge base comprises three major knowledge levels, consisting of several independent but interrelated modules:

Lexical level:

- The Lexicon stores morphosyntactic, pragmatic and collocational information about lexical units.
- The Morphicon helps our system to handle cases of inflectional morphology.

Grammatical level:

- The Grammaticon is composed of several Constructicon modules[3] whose constructional schemata help Role and Reference Grammar (RRG) to build the semantics-syntax-semantics linkage (Van Valin and LaPolla, 1997; Van Valin, 2005).

Conceptual level:

- The Ontology is presented as a hierarchical catalogue of the concepts that a person has in mind, so here is where semantic knowledge is stored in the form of meaning postulates. The Ontology consists of a general-purpose module (i.e. Core Ontology) and several domain-specific terminological modules (i.e. Satellite Ontologies).
- The Cognicon stores procedural knowledge by means of scripts, i.e. conceptual schemata in which a sequence of stereotypical actions is organised on the basis of temporal continuity, and more particularly on the basis of Allen's temporal model (Allen, 1983; Allen and Ferguson, 1994).
- The Onomasticon stores information about instances of entities and events, such as Bill Gates or 9/11. This module stores two different types of schemata (i.e. snapshots and stories), since instances can be portrayed synchronically or diachronically.

In the FunGramKB architecture, every lexical or grammatical module is language-dependent, whereas every conceptual module is shared by all languages. In other words, computational linguists must develop one Lexicon, one Morphicon and one Grammaticon for English, one Lexicon, one Morphicon and one Grammaticon for Spanish and so on, but knowledge engineers build just one Ontology, one Cognicon and one Onomasticon to process any language input conceptually. In this scenario, the Ontology becomes the pivotal module for the whole architecture.

3 *UniArab*

UniArab—Universal Arabic Machine Translator (Salem, Hensman and Nolan, 2008a, 2008b; Nolan and Salem, 2009; Salem and Nolan, 2009a, 2009b) is able to provide a working translation of Modern Standard Arabic[4] to English. UniArab covers a representative broad selection of words and can translate simple sentences including intransitive, transitive and ditransitive clauses, as well as copular-like nominative clauses.

UniArab is built upon an interlingua machine translation architecture, which is more flexible and scalable for multilingual generation. Indeed, an accurate representation of the RRG logical structure of an Arabic sentence is one of the primary strengths of UniArab. RRG is one of the most relevant functional models on the linguistic scene today. This grammatical model adopts a communication-and-cognition view of language, i.e. morphosyntactic structures and grammatical rules should be explained in relation to their semantic and communicative

[2] English and Spanish are fully supported in the current version of FunGramKB, although we have just begun to work with other languages, i.e. German, French, Italian, Bulgarian, Catalan and Arabic.

[3] The Grammaticon distinguishes four levels of meaning construction: argumental layer (L1-Constructicon), implicational layer (L2-Constructicon), illocutionary layer (L3-Constructicon), and discoursive layer (L4-Constructicon).

[4] Modern Standard Arabic is the literary and standard variety of Arabic used in writing and formal speeches today (Schulz 2005).

functions. In RRG, the semantic and the syntactic components are directly mapped in terms of a linking algorithm, which includes a set of rules that account for the syntax-semantics interface. As a result, RRG allows an input text to be represented in terms of a logical structure, which has been enhanced by a new formalism called “conceptual logical structure” (Periñán-Pascual and Mairal-Usón, 2009).

Concerning the evaluation of UniArab, the results obtained by this machine translator were compared to their human-translated equivalents and to the results obtained by the translation services from Google (2009) and Microsoft (2009). Although these statistical machine translators have wider coverage, Salem and Nolan (2009b) demonstrated that UniArab provides more accurate and grammatically-correct translations.

However, the model of UniArab devised by Brian Nolan and his research team fails to provide an adequate treatment of the morphology and the semantics of lexical units. On the one hand, the UniArab Lexicon stores a separate lexical entry for every allomorph of a lemma, so inflectional morphology is not handled efficiently. On the other hand, UniArab avoids the problem of word sense disambiguation by adopting a naive one-word-one-sense approach to lexical polysemy. To overcome these deficiencies, one of our objectives is to replace the UniArab lexical database by a robust knowledge base such as FunGramKB.

4 Integrating FunGramKB into UniArab

The new architecture of UniArab, in which FunGramKB has been fully integrated, breaks down into seven tasks: (1) tokenization, (2) morphological parsing, (3) syntax-semantics processing, (4) CLS construction, (5) syntactic generation, (6) morphological generation and (7) sentence generation. A brief description of these tasks is presented in the remainder of this section.

4.1 Task 1: tokenization

The input is split into sentence tokens, and then into word tokens. In other words, the first task consists in segmenting the input into basic units of analysis, which mostly correspond to orthographic words. This phase also involves some pre-processing, identifying what is not to be translated, such as punctuation marks and symbols.

4.2 Task 2: morphological parsing

Lexical tokens are deprived of their inflectional affixes. The FunGramKB Lexicon adopts a lemma-based model, since regular inflected forms can be generated by a morphological component provided with inflectional patterns in the form of rules. More particularly, cases of inflectional morphology are handled by the FunGramKB Morphicon, which is made up of two submodules: MorphoRules, which contains a set of regular expression rules, and MorphoDB, a database of irregular word-forms.

4.3 Task 3: syntax-semantics processing

In this phase, the system resolves word sense disambiguation and determines the phrasal structure of the input, being both tasks performed in a parallel fashion. According to Mahesh (1995), neither sequential architectures, where a lower-level process does not get any feedback from a higher-level process, nor integrated architectures, where the different types of knowledge are not applied independently, are appropriate to model the syntax-semantics interaction in sentence understanding. In the new architecture of UniArab, a parallel configuration with an interactive controller preserves the independence of syntax and semantics while permitting bi-directional communication between the lexico-semantic parser and the syntactic parser.

At the end of this task, the system generates a syntactic representation of the input where lemmas have been replaced by conceptual tags. To illustrate, suppose that we want to translate into English the Arabic sentence (1), whose translation equivalent is (2).

(1) قرأ خالد الكتاب

(2) Khalid read the book.

In this case, the syntax-semantics processing outputs the parenthetical representation (3).

(3) S(NP(n(%KHALID_00)),
VP(v(+READ_00), NP(det(the),
n(+BOOK_00))))

Together with the representation (3), the system should also hold all features and values specific to each content word (i.e. adjective, noun or verb) in the input. For example, the feature-value structures (4) and (5) show the information assigned to the words قرأ (i.e. read—past tense) and يقرأ (i.e. read—infinitive) according to the old and new models of UniArab respectively.

(4)

a- Source word:	قرأ
b- Category:	v
c- Gender:	m
d- Number:	sing
e- Person:	3rd
f- Tense:	past
g- Logical structure:	<TNS: PAST [do(x,[read(x,y)])]>
h- Translation (ENG):	read

(5)

a- Source word:	يقرأ
b- Concept:	+READ_00
c- Category:	v
d- Gender:	m
e- Number:	sing
f- Person:	3rd
g- Tense:	past
h- Aktionsart:	Active accomplishment

i- Thematic-frame mapping:	x = Theme
	y = Referent

j- Translation (ENG):	read

Comparing both structures, we conclude that the difference does not lie just in the type and amount of information generated, but also in the way this information is obtained. In the case of (4), all the values are retrieved from the lexical entry, since a wordform-based model of computational lexicon is used. The advantage of this type of model deals with the simplicity to parse the input. However, this approach presents some drawbacks: redundancy of information, inefficient management of lexica and inability to predict new inflected forms (Lehrberger and Bourbeau, 1988; Trost, 2003). On the contrary, in the case of (5), values from features (d-g) are generated by the morphological parser. However, the most remarkable difference can be found in the logical structure, which is underspecified in features (h) and (i) in (5). The motivation of this approach is described in the task 4.

Since lexical information in FunGramKB is linked to the senses of words (i.e. sense-oriented approach), a word-sense disambiguator should firstly tag the lemmas with a single conceptual label from the Ontology. Lexical ambiguity takes place when a word is linked to more than one concept in the Ontology. Although the semantic polyvalence of words contributes to the economy of language, this phenomenon poses a serious problem in machine translation. Unlike the original model of UniArab, our approach resolves lexical ambiguity by allowing the machine to obtain the most suitable type of information out of the context of the target word. For this purpose, Higinbotham's algorithm (1990: 134-144) for word sense disambiguation has been adapted to the characteristics of FunGramKB, as shown in Figure 1.

For instance, if the sentence (2) were taken as the input text, the only ambiguous word would be *book*, since it is the headword of two lexical entries: (i) a set of printed pages and (ii) to make a reservation. According to Figure 1, this word would be easily disambiguated on the basis of its part of speech.

The word-sense disambiguation algorithm controls the access to various types of knowledge in FunGramKB, mainly from the Lexicon, the Grammaticon and the Ontology:

- Lexicon: idioms, morphosyntactic constraints (e.g. part of speech, number, gender, countability, adjectival position, pronominalization, etc), collocations, domain, and frequency.
- Grammaticon: high-level constructional schemata.
- Ontology: selectional preferences in the thematic frame, and spreading activation taking the form of the MicroKnowing.

The most complex task involved in this algorithm is our spreading activation method. The rationale behind the spreading activation is that the neighbouring context of the ambiguous word can trigger connections with many meaning representations in the knowledge base. In FunGramKB, the spreading activation is performed by the MicroKnowing, i.e. Microconceptual-Knowledge Spreading (Periñán-Pascual and Arcas-Túnez, 2005), which can be defined as a multi-level pre-reasoning process for the construction of the extended meaning postulates of FunGramKB concepts.

```
IF the word is linked to more than one meaning THEN
    IF the word occurs in an idiom or in a high-level constructional scheme THEN
        Translate the entire phrase with its idiomatic meaning
    ELSE
        Check morphosyntactic constraints
        IF only one meaning is left THEN
            Use that meaning
        ELSE
            Check selectional preferences
            IF only one meaning is left THEN
                Use that meaning
            ELSE
                Search for previous occurrences
                IF the word has already occurred in the text THEN
                    Use the same meaning used in the previous occurrence
                ELSE
                    Check technical domains
                    IF there is one matching with the input domain THEN
                        Use that meaning
                    ELSE
                        Apply a spreading activation method
                        IF there is a winning candidate THEN
                            Use that meaning
                        ELSE
                            Take the most frequent meaning
                        END IF
                    END IF
                END IF
            END IF
        END IF
    END IF
END IF
```

Figure 1: Word-sense disambiguation algorithm.

One of the benefits of the MicroKnowing for a natural language processing knowledge base such as FunGramKB is that redundancy is minimized, whereas informativeness is maximized, in the semantic and common-sense knowledge repository. Therefore, a meaning postulate which is stored in FunGramKB does not necessarily include all the generic features of the *definiendum*, but just those predications which cannot be retrieved anywhere else in the knowledge base. For instance, the predications in the meaning postulate of +BIRD_00 are presented in (6).

(6) +(e$_1$: +BE_00 (x$_1$: +BIRD_00)$_{\text{Theme}}$ (x$_2$: +VERTEBRATE_00)$_{\text{Referent}}$)
*(e$_2$: +COMPRISE_00 (x$_1$)$_{\text{Theme}}$ (x$_3$: $_m$ +FEATHER_00 & $_2$ +LEG_00 & $_2$ +WING_00)$_{\text{Referent}}$)
*(e$_3$: +FLY_00 (x$_1$)$_{\text{Agent}}$ (x$_1$)$_{\text{Theme}}$ (x$_4$)$_{\text{Origin}}$ (x$_5$)$_{\text{Goal}}$)

In this example, the MicroKnowing allows the system to enrich the conceptual representation of +BIRD_00 by retrieving other prototypical features such as “birds lay eggs” or “birds can breathe” from the meaning postulates linked to other concepts in the Ontology. Computationally speaking, this is feasible because the MicroKnowing takes place in a multi-level scenario, since it is performed by the iterative application of two types of reasoning mechanisms: inheritance and inference. Whereas inheritance strictly involves the transfer of one or more predications from a superordinate concept to a subordinate one in the Ontology, our inference mechanism is based on the constructs shared between predications linked to conceptual units which do not take part in the same subsumption relation within the Ontology. In this way, the MicroKnowing allows the system to reveal the semantic and common-sense knowledge underlying the input text. Finally, a measure of relatedness such as the Adapted Lesk Algorithm (Banerjee and Pedersen, 2002; Patwardhan, Banerjee and Pedersen, 2003) should be applied to the inventory of extended meaning postulates resulting from the MicroKnowing triggered by the nouns and verbs present in the input. The winning candidate is that meaning of the ambiguous word with the highest index of relatedness with respect to the meanings of the neighbouring words in the source text.

4.4 Task 4: CLS construction

The conceptual logical structure (CLS) is developed out of the phrasal structure together with the lexico-conceptual information obtained in the task 3. The creation of the CLS is the most crucial phase along this translation process.

As can be seen in (4), the whole logical structure of the lexical unit *read* was stored in the lexical entry. On the contrary, the FunGramKB CLS Constructor can build automatically the representation (7) from the *Aktionsart* and *Thematic-frame mapping* features in the Lexicon and from the conceptual knowledge stored in the Ontology, together with the application of the RRG linking algorithm.

(7) do (x_{Theme}, [+READ_00 (x_{Theme},
($y_{Referent}$))] & INGR +READ_00
($y_{Referent}$)

This shift from the standard RRG model of logical structure to the CLS approach makes the system handle real language-independent representations, since these are made of concepts and not words.

Following the example (1), the output of this task can be seen in the CLS (8), whose conceptual instanciations are also enriched by lexico-conceptual information represented in the form of feature-value matrices.

(8) $<_{IF}$ DECL $<_{TNS}$ PAST < do
(%KHALID_00$_{Theme}$, [+READ_00
(%KHALID_00$_{Theme}$,
+BOOK_00$_{Referent}$)] & INGR
+READ_00 (+BOOK_00$_{Referent}$)>>>

4.5 Task 5: syntactic generation

The English Grammaticon maps the CLS into a structured syntactic representation on the basis of the RRG linking algorithm. Thus, the syntactic representation for the CLS (8) is as follows:

(9) S(NP(n(Khalid)), VP(v(read),
NP(det(the), n(book))))

As can be noted, concepts have been now replaced by the translation equivalents of the source words.

4.6 Task 6: morphological generation

In this phase, lemmas in the syntactic representation are replaced by suitable word-forms. This requires that the feature-value structures should be accommodated to the information of the target words, so the target Lexicon plays an important role in this task.

Although FunGramKB relies heavily on the Morphicon for the construction of inflectional forms, some lexical features of the target words help the system to determine the morphological submodule to be triggered: MorphoRules or MorphoDB. For example, the lexical entry of *read* states that its verb paradigm is irregular, so expression rules from MorphoRules are blocked out in benefit of ready-made morphological forms from MorphoDB.

4.7 Task 7: sentence generation

Finally, the output of the sentence generator takes the form of sentence (2). The architecture of the enhanced UniArab system is shown in Figure 2.

5 Conclusions

The advantage of UniArab lies in the deployment of an interlingua architecture which uses a robust functional linguistic model in the machine translation kernel. UniArab clearly outperforms existing systems in the processing of simple sentences, suggesting that RRG is a promising candidate for Arabic-to-English machine translation.

By replacing UniArab lexical database by FunGramKB, where lexical entries are more informative and meaning capabilities are deeper, the system can begin to cope with complex multilingual input.

Acknowledgement

Financial support for this research has been provided by the DGI, Spanish Ministry of Education and Science, grant FFI2008-05035-C02-01/FILO. The research has been co-financed through FEDER funds.

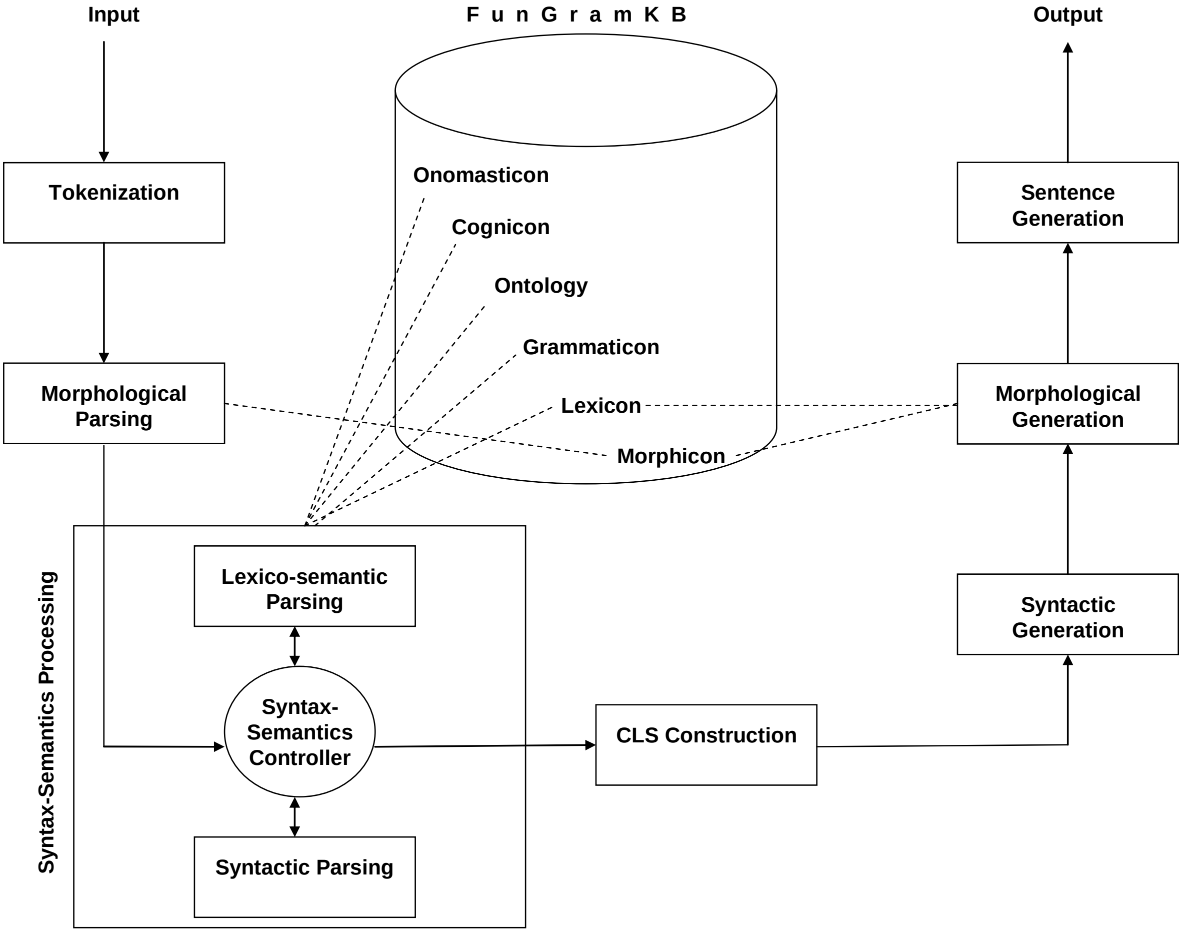

Figure 2: The new architecture of UniArab.

References

Allen, J.F. 1983. Maintaining knowledge about temporal intervals. *Communications of the ACM*, 26 (11): 832-843.

Allen, J.F. and G. Ferguson. 1994. Actions and events in interval temporal logic. *Journal of Logic and Computation*, 4 (5): 531-579.

Banerjee, S. and T. Pedersen. 2002. An adapted Lesk algorithm for word sense disambiguation using WordNet. In *Proceedings of the 3rd International Conference on Computational Linguistics and Intelligent Text Processing*. Springer, London, 136-145.

Google. 2009. Google Translator. http://translate.google.com.

Higinbotham, D.W. 1990. *Semantic cooccurrence networks and the automatic resolution of lexical ambiguity in machine translation*. Ph.D. thesis. University of Texas, Austin.

Lehrberger, J. and L. Bourbeau. 1988. *Machine Translation: Linguistic Characteristics of MT Systems and General Methodology of Evaluation*. John Benjamins, Amsterdam: Philadelphia.

Mahesh, K. 1995. *Syntax-semantics Interaction in Sentence Understanding*. Ph.D. thesis. Georgia Institute of Technology, Atlanta.

Mairal Usón, R. and C. Periñán Pascual. 2009. The anatomy of the lexicon component within the framework of a conceptual knowledge base. *Revista Española de Lingüística Aplicada*, 22: 217-244.

Microsoft. 2009. Microsoft Translator. htttp://www.windowslivetranslator.com/Default.aspx.

Nolan, B. and Y. Salem. 2009. UniArab: an RRG Arabic-to-English machine translation software. In *Proceedings of the Role and reference Grammar International Conference*. University of California, Berkeley.

Patwardhan, S., S. Banerjee and T. Pedersen. 2003. Using measures of semantic relatedness for word sense disambiguation. In *Proceedings of the 4th International Conference on Intelligent Text Processing and Computational Linguistics*. Ciudad de Méjico, 241-257.

Periñán-Pascual, C. and F. Arcas-Túnez. 2004. Meaning postulates in a lexico-conceptual knowledge base. In *Proceedings of the 15th International Workshop on Databases and Expert Systems Applications*. IEEE, Los Alamitos (California), 38-42.

Periñán-Pascual, C. and F. Arcas-Túnez. 2005. Microconceptual-Knowledge Spreading in FunGramKB. In *Proceedings of the 9th IASTED International Conference on Artificial Intelligence and Soft Computing.* ACTA Press, Anaheim-Calgary-Zurich, 239-244.

Periñán-Pascual, C. and F. Arcas-Túnez. 2007. Cognitive modules of an NLP knowledge base for language understanding. *Procesamiento del Lenguaje Natural,* 39: 197-204.

Periñán-Pascual, C. and R. Mairal Usón. 2009. Bringing Role and Reference Grammar to natural language understanding. *Procesamiento del Lenguaje Natural,* 43: 265-273.

Salem, Y., A. Hensman and B. Nolan. 2008a. Implementing Arabic-to-English machine translation using the Role and Reference Grammar linguistic model. In *Proceedings of the 8th Annual International Conference on Information Technology and Telecommunication,* Galway, Ireland.

Salem, Y., A. Hensman and B. Nolan. 2008b. Towards Arabic to English machine translation. *ITB Journal,* 17: 20-31.

Salem, Y. and B. Nolan. 2009a. Designing an XML lexicon architecture for Arabic machine translation based on Role and Reference Grammar. In *Proceedings of the 2nd International Conference on Arabic Language Resources and Tools,* Cairo, Egypt.

Salem, Y. and B. Nolan. 2009b. UniArab: a universal machine translator system for Arabic Based on Role and Reference Grammar. In *Proceedings of the 31st Annual Meeting of the Linguistics Association of Germany*.

Schulz, E. 2005. *A Student Grammar of Modern Standard Arabic*. Cambridge University Press, Cambridge.

Trost, H. 2003. Morphology. In R. Mitkov, ed. 2003. *The Oxford Handbook of Computational Linguistics*. Oxford University Press, Oxford, 25-47.

Van Valin, R. 2005. *Exploring the Syntax-Semantic Interface*. Cambridge University Press, Cambridge.

Van Valin, R. and R. LaPolla. 1997. *Syntax: Structure, Meaning, and Function*. Cambridge University Press, Cambridge.