

Transducing Arabic Phrases into Head-Modifier (HM) Pairs for Arabic Information Retrieval

W. Everhard Ditters, Cornelis H.A. Koster

Humanities Department, Computer Science Institute

University of Nijmegen

The Netherlands

e.ditters@let.kun.nl, kees@cs.kun.nl

0 Abstract

In order to raise the precision of Information Retrieval systems, linguistically derived phrases may be used besides the traditional (single) keywords for both the document representation and the queries. Such phrases may either take the form of (monolithical) collocations or, as shown in this paper, of Head/Modifier pairs representing dependency structures in the text.

In this paper we describe the rationale behind the Head/Modifier approach, which has been developed for English, and investigate its applicability to Arabic texts. In particular, we show how it has to be adapted in order to cope with the rather distinct syntactic particularities of the languages involved.

Keywords: Natural Language Processing, Information Retrieval, Arabic, English, Head/Modifier pairs.

1 Introduction

The representation of a document by a bag-of-words, as traditionally used in Information Retrieval, is repugnant to any person with linguistic leanings, because it obviously throws away important information contained in the structure and ordering of phrases. That is why the use of Linguistically Motivated Indexing terms (some abstraction of linguistically derived phrases) has always fascinated researchers in IR (for an overview see Sparck Jones 1999).

As a search term, a phrase like “software engineering” is obviously much more precise than its constituent words, therefore it makes sense to at least extend the keyword approach with phrases. In particular the use of simple noun phrases as indexing terms is now commonly accepted, but the use of phrases with verbs is still a challenge.

Two approaches can be distinguished: to employ a linguistically motivated collocation or multi-word unit as a single monolithical term like “software_engineering” (e.g. Lewis 1992, Smeaton 1997) or to use as terms pairs consisting of a head and a modifier, like “[engineering, software]” (e.g. Fagan 1988, Strzalkowski 1995, Bolshakov 2004). The modifier serves to precise and disambiguate the head, and vice versa.

Our approach (described in Koster 2004) is based on Head/Modifier pairs. We have built a system for the robust recognition of phrases in running English text, their transduction to Head/Modifier trees (binary dependency trees) and the un-nesting of those dependency trees to Head/Modifier pairs (see www.cs.kun.nl/agfl). Central to this system is a grammar for English developed for IR applications (EP4IR) which was written in the AGFL formalism (Koster 1991) and from which the parser is automatically generated.

In the course of the parsing and transduction process, the system also performs the robust recognition of Named Entities and out-of-vocabulary word forms, as well as a number of syntactic normalizations which serve to increase recall without impairing precision.

In the present paper we investigate the applicability of Head/Modifier approach to Modern Standard Arabic, as part of the development of a similar grammar (ARP4IR) for Information Retrieval applications.

Modern research in automatic linguistic analysis, like research in Information Retrieval, is predominantly concerned with English text data, with at best some sidesteps towards carefully selected illustrative examples from “exotic” languages. In our approach we try to profit from the available English experience while avoiding language dependent pitfalls in applying the same computational framework to Arabic language data.

In the rest of the paper we will first briefly speak about the rationale for Head/Modifier pairs in the context of English (§ 2) and the automatic syntactic analysis of Arabic text data (§ 3). In § 4 we will list the building blocks for information retrieval from Arabic text data. In the next section we briefly speak about the IR-tools for Arabic text data: the parser, the lexicon and the test corpus (§ 5). In the meantime, parallels will be drawn with results from the use of the same framework for IR of English text data. Finally we will end with a conclusion (§ 6) and references (§ 7).

2 Head-Modifier Pairs in English

In Information Retrieval, many document representations have already been tried besides the traditional bag-of-words

representation, consisting of bags or sets of: n-grams of characters, n-grams of words, lemmatized or stemmed words, chunks, collocations and phrases, which were obtained by statistical or syntactical means. For English and other languages with little inflection, the bag-of-words representation still outperforms all others, but for highly inflected languages like Arabic more linguistics is necessary (at least some form of stemming).

Although the use of simple noun phrases as indexing terms is now commonly accepted, practical Information Retrieval systems using phrases like the Clarit system (Evans et al, 1993) do not appear to perform consistently better than those based on keywords. The reason for this is found in the statistical distribution of phrases, of which there are very many more than of words, and which occur much more sparsely.

Although phrases may be much more precise than single words, their recall is much smaller. Recall must be painfully regained by normalizing transformations, mapping for as far as possible all equivalent forms of a phrase onto one same term by means of morphological normalization (stemming), syntactical normalization and semantical normalization (Arampatzis et al, 1998), thus undoing the effects of linguistic variation.

The ideal document representation captures the ‘aboutness’ of the document (see Bruza and Huijbers, 1996) without retaining useless features of the document. In Arampatzis et al (2000) it was found that in the context of document categorization the nouns, adjectives and verb forms in an English document carry nearly all of its ‘aboutness’, and that all other words may safely be discarded. In particular, auxiliary verbs indicating time and modality as well as adverbs were found to be redundant. Therefore in Koster (2004) it was argued that in a Head/Modifier pairs representation only those pairs of which the head or the modifier is a noun, adjective or verb form carry the ‘aboutness’. Pronouns are also included for possible anaphora resolution. This leads to the following taxonomy of HM pairs for the English language (in which V stands for verb, N for noun, P for (personal) pronoun, A for adjective and PP for a preposition phrase:

I	Modifier				
	V	N	P	A	PP
V	-	object relation		-	+
N, P	subject relation	predicative/attributive relation			+
A	-	-	-	-	+

Figure 1: Relations between HM pairs

We next show some examples of phrases with the corresponding HM pairs, for each of the four realizations in the above taxonomy. Each pair is derived from the dependency tree by un-nesting and removal of redundant pairs. Note that, except for the PP which is marked by its preposition, the elements of a pair are all typed as a noun (N:), verb (V:), Adjective (A:), or pronoun (P:) and have been lemmatized accordingly.

- a) The Subject relation (NV)
- | | |
|------------------|------------------|
| this man sneezed | [N:man,V:sneeze] |
| the sneezing man | [N:man,V:sneeze] |
| I walk | [P:I,V:walk] |
| I like to walk | [P:I,V:walk] |

Although a sentence like “I like to walk” might suggest a pair [V:like,V:walk], this sentence has in fact the same aboutness as “I walk”. We therefore prefer the above transduction over [P,I,V:like][V:like,V:walk]. The status of VV pairs in English is dubious.

- b) The Attributive and Predicative relation are not distinguished (NA and NN)
- | | |
|----------------------|----------------------------|
| this car is red | [N:car,A:red] |
| my red car | [N:car,A:red] |
| software engineering | [N:engineering,N:software] |
- c) The Object relation occurs with transitive verbs and participles (VN)
- | | |
|------------------------------------|---|
| I hate you | [P:I,V:hate][V:hate,P:you] |
| it could have hit me | [P:it,V:hit][V:hit,P:me] |
| IBM sponsored this conference | [N:IBM,V:sponsor][V:sponsor,N:conference] |
| the conference is sponsored by IBM | [N:IBM,V:sponsor][V:sponsor,N:conference] |
| a sponsored conference | [P:it,V:sponsor][V:sponsor,N:conference] |

Notice that in the last example the participle is not transduced as an adjective, but as a verb with a filler (for anaphora) as subject.

- d) The PP relation has a preposition marking the modifier, and can occur with N or P, A and V heads.
- | | |
|--------------------|---|
| I gave you a knife | [P:I,V:give][V:give,N:knife][V:give,to P:you] |
| a cry for freedom | [N:cry,for N:freedom] |
| open to suspicion | [A:open,to N:suspicion] |

In the rest of this article, we shall try to derive the corresponding taxonomy for Arabic.

3 Building blocks for the linguistic analysis of Arabic text data

For analysis purposes we apply a linguistic description in terms of immediate constituents enhanced with a component to account for relationships and dependencies between the constituents of a sentence as well as between the elements of a constituent. This two-level linguistic description, formalized within the conditions of Attribute or Affix Grammars (Koster, 1991), goes top-down, from abstract to specific, alternating between functional and categorical layers until the final lexical entries have been reached. Keywords on this level thus are: syntactic structure, dependency structure, normalized recursivity, function and category, head and modifier, affixes, affix domains and affix

values (the term affix here to be understood as variable in a formal description with a finite domain of values).

Theoreticians of the early Arabic grammatical tradition adopted a distinction of parts of speech into: the category of noun (**N**), verb (**V**) and particle (**P**) or rather ‘that what’ is neither noun nor verb ($\neg N|V = P$).¹ The matrix in figure 2 tells us about the possible combinations of the elements in column 1 with the elements of row 1 in order to fill in functional slots at the next level of description.

2	Modifier			
Head	N	V		P
N	+	+		+
V	+	+		+
P	+	+		+

Fig. 2: Possible **HM** combinations

Interpreting the Arabic grammatical tradition Wright (1974², vol.I, p.278) lodged conjunctions, interjections, prepositions and adverbs in the **P** category, while, on the contrary, El-Ayoubi et al. (2003, T1, B2, p. 462), place (most of) the prepositions and (most of) the adverbs on formal grounds in the category **N**.

We prefer to follow the early Arabic grammatical tradition but we expand the two head elements **N** and **V** into the phrasal categories **NP** and **VP** in order to account for possible extensions or modifiers of these head elements with other NP’s, VP’s, adjective phrases (**ADJP**), adverb phrases (**ADVP**), prepositional phrases (**PP**) and different types of clauses (**CL**) such as prepositional clauses (**PCL**), complement clauses (**CCL**), relative clauses (**RELCL**) and conditional clauses (**CONCL**).⁴ These phrasal categories fill in functional slots realizing one of the two basic types of the Arabic language sentence structure: the nominal (**Sⁿ**) and the verbal (**S^v**) sentence, or the functional slots at lower levels of the linguistic description.

3	Modifier		
Head	N	V	P
N	NP S ⁿ	NP S ⁿ	S ⁿ
V	VP (REL)CL S ^v	VP CL S ^v	CL S ^v
P	PP	PCL	ADVP

Fig. 3: syntactic realization of **HM** pairs

So far we spoke in terms of categories. As a matter of fact we distinguish between obligatory and optional functions and categories (Ditters, 2001) realizing these functions at different levels starting with the sentence as, for the moment, our highest unit of linguistic description as shown in figure 4.

4	Sentence level					
S-type	Functions	Categories ⁵				
S ⁿ CL ⁿ	topic			CCL ⁶	NP	
	comment	ADJP	ADVP	CCL	NP	VP
	[sentence adverbial]		ADVP	CCL	NP	PP
S ^v CL ^v	predication					VP
	[sentence adverbial]		ADVP	CCL	NP	PP

Fig. 4: functions and categories at sentence level

The functional structure map of the categories playing a role at sentence level are shown in the following figure 5.

5	Phrase level						
Category	Function ³						
ADJP		[NEG]	[PREDET]		HEAD	[POSTDET]	[POM]
ADVP					HEAD		[POM]
CCL				LINKER			CL-COMPL
NP		[NEG]	[PREDET]		HEAD	[POSTDET]	[COMPL]
PP PCL				LINKER			(P PCL)-COMPL
VP	[PREM]	[NEG]			HEAD		[COMPL]

¹ We use the vertical bar as separator for alternatives and the not-sign as negation operator.

² This date refers to the new impression. As a matter of fact, this new impression is Cachia’s slightly annotated version of de Goeije’s 3rd edition of William Wright’s revised and enlarged 2nd edition of his translation of Caspari’s Arabic Grammar and published the 1st of July, 1874.

³ The different functions under Phrase level are ordered sequentially.

⁴ In the following figures parentheses mark possible realizations and square brackets optional functions.

⁵ The different categories under Sentence level are ordered alphabetically.

⁶ The complement clause (CCL) comprises: finite as well as non-finite clauses, that-clauses, prepositional clauses, conjunctive clauses like by ‘wa’-introduced ‘hal’-sentences. However, English ‘to’-clauses are NP’s in Arabic.

For the analysis process we consider to be ‘head’ of a constituent the first element marked for its function at the next-higher level of description. Moreover, we assume that a phrasal head usually can be identified by a single syntactic form with a semantic value (coinciding phrasal heads). In a number of occurrences we find distinct head realizations of the syntactic form and the semantic value within a constituent. Hence we spoke about non-coinciding phrasal heads within a constituent (Ditters, 2003a and b; cfr. El-Ayoubi et al, 2001, T1, B1, p. 120). In this way ‘pre-posed’ demonstratives, quantifiers, modal verbs and auxiliaries are considered to be the syntactic head of a phrase governing a second element being the semantic head of the verb cluster.

4 Building Blocks for the Information Retrieval from Arabic Text Data

In Figure 5 we listed the head/(post)modifier combinations for automatic analysis purposes at phrase level. For IR-purposes we will only take into consideration head/modifier and linker/complement combinations at this level with a special emphasis on the NP and the VP. As for the head/modifier combinations of the ADJP, ADVP and the linker/complement combination of the PP, they are considered to be of basic importance when realizing obligatory functions at sentence level and only of additional or supplementary, let us say negligible, importance as far as they are instantiated as optional (post)modifying elements within the NP, the VP or other phrasal categories.

We will use the HEAD, [POSTDET], [POM], [COMPL] and the HEAD, [POM], [COMPL] combinations, respectively: NP’s and VP’s in Arabic text data. Moreover, for IR-purposes it seems not to be productive to maintain a distinct subclass of the N category in the form of denominal or deverbal adjectives (A) because of a minor difference in structural combinatorial behaviour when occurring as (post)modifier in a NP. We still consider ADJP’s as a subclass of the N-category with a limited attributive and/or predicative function at phrase level.

Interestingly enough, the classical Arabic paradigm of parts of speech answers to the needs for robust information retrieval (IR) of Arabic documents. We are interested in the N–N, N–V, V–N and V–V combinations as well as their obligatory or optional extensions with the non-Noun (\neg N) and non-Verb (\neg V) category (\neg N|V = P) exclusively used as fillers in the function of modifier. Copying the scheme of HM-pairs for English in figure 1 we may obtain for Arabic:

6	Modifier					
	Head	V	N	P	A	P
V	inversed relation	subject/object relation			-	+
N, P	topic/comment relation	Possessive /predicative / attributive relation				+
A	-	+	-	-	+	

Fig.6: possible HM-pairs for Arabic

For IR-purposes we disregard any sentence adverbial at the sentence level. At phrasal level we disregard any predetermining and premodifying element as well as phrasal adverbials. At morpho-syntactic level we disregard for the verbs the affix values for ‘aspect’, ‘mode’ and ‘voice’. On the contrary, we maintain the information concerning ‘stem’ and ‘complementation’ with information about the complement structure to be matched in the lexicon. As far as the noun is concerned, we disregard affix values concerning ‘gender’, ‘number’ and ‘case’ while broken plurals are matched via the lexicon into the main category N. For the subclass pronouns we maintain the affix ‘person’ while disregarding specific affix values.

We next show some examples of phrases with HM pairs involving elements of the N and V categories. The pairs we would like to deal with here are the: NN, the NV, the VN and the VV. Each pair is derived from the dependency tree by the un-nesting and removal of redundant pairs. Note that the elements of a pair all typed as a noun (N:), a verb (V:) and, occasionally, an adjective (A:) and a prepositional phrase (PP:) have been lemmatized accordingly

In the examples we first present the transcribed Arabic data, a morph-to-morph translation in English, followed by the original translation, doubled by an English translation and the source of the quotation.

NN : The NN pair constitutes at sentence level a Sⁿ expressing a topic-comment (= Head-Modifier) relation. The modifier may belong to the category noun (example 1) or can be an adjective, a subclass of the nouns (example 2). In both cases we are dealing with a predicative relation between the modifier to the head.

- (1) hādā ’abū bakrīn
this Abu Bakr
This is Abu Bakr (Cant.:I,18,4)⁷
[N: hādā, N: ’abū bakr]
- (2) ’alsamā’u ḡamīlatun
the-sky beautiful
The sky is beautiful (Cant.:I,15,1)
[N: samā’, N: ḡamīl]

The NN pair constitutes at phrasal level an NP expressing a head-modifier relation. Both the NN (example 3) and the NA realization (example 4) may express an appositive relation.

- (3) ’alraḡulu ’lḡarību māliku ’lṣayfi
the-man the-strange owner the-sword
der fremde Mann, der Besitzer des Schwerts
(Ayoubi:I,I,453,2)⁸

⁷ The source indication refers to: Cantarino, Volume x, page y, example z.

The strange man in possession of the sword.
 [N: rağul, N: ġarīb] [N: rağul, A: mālik]
 [A: mālik, N: sayf]

- (4) 'alṭā'iru 'lmaksūru 'lġanāḥayni
 the-bird the-broken the-wings
 The bird with broken wings (Cant.:II,110,8)
 [N: ṭā'ir, A: maksūr] [A: maksūr, N: ġanāḥ]

Moreover, an NN pair may express a partitive (example 5), possessive (example 6), a subject (example 7) or an object relation (example 8) between the head and the modifier.

- (5) zaytu 'lzaytūni
 oil the-olive
 das Öl der Oliven (Ayoubi:I,I,497,4)
 Olive oil
 [N: zayt, N: zaytūn]
- (6) kitābu ṭālibin
 book student
 (das) Buch eines Studenten (Ayoubi:I,I,498,2)
 The book of a student
 [N: kitāb, N: ṭālib]
- (7) ba'ada ruġū'iy
 after return-mine
 After my return (Cant.:II,402,3)
 [PP: ba'ada, N: ruġū'] [N: ruġū', N: 'anā⁹]
- (8) bi'idḥāli yadihi
 by-introduction hand-his
 to introduce his hand (Cant.:II,402,7)
 [PP: bi, N: 'idḥāl] [N: 'idḥāl, N: yad]
 [N: yad, N: huwa]

NV : The NV pair constitutes at sentence level a Sⁿ expressing a topic-comment (= Head-Modifier) relation (example 9). At phrasal level the NN pair constitutes an NP expressing a head-modifier relation¹⁰ (example 10).

- (9) mabādi'u zālat
 principles disappeared-she
 Principles disappeared (Cant.:I,93,1)
 [N: mabādi', V: zāla] [V: zāla, N: hiya]
- (10) malā'ikatun nazalū mina 'lsamā'i
 angels came-down from the-heaven
 Angels who have come down from heaven
 (Cant.:I,93,4)

[N: malā'ikat, V: nazala] [V: nazala, N: huwa]
 [V: nazala, min, N: samā']

VN : The VN pair constitutes at sentence level a S^v realizing the predicate function (= Head-Modifier relation) (example 11). At phrasal level the VN pair constitutes a VP expressing a head-modifier relation (see example 10 above) in which the modifier may realize the subject (example 12), the direct (example 13), the prepositional object (example 10) or the benefactive relation (example 14).¹¹

- (11) daḥala 'lnabiyyu 'ilay baytihi
 went-he the-prophet in house-his
 The prophet went into the house (Cant.:I,45,9)
 [V: daḥala, N: nabiyy]
 [V: daḥala, 'ilay, N: bayt] [N: bayt, N: huwa]
- (12) tataġāwabu 'aṣḍiqā'uhā
 reply-he friends-her
 Her friends reply (Cant.:I,87,2)
 [V: taġāwaba, N: ṣadiq] [N: ṣadiq, N: hiya]
- (13) daḥala 'lmaḍinata
 entered-he the-city
 He entered the city (Cant.:II,163,1)
 [V: daḥala, N: huwa] [V: daḥala, N: maḍinat]
- (14) nāwalaniy 'iyyāhu
 handed-he-me it
 He handed it to me (Cant.:II,169,1)
 [V: nāwala, N: huwa] [V: nāwala, N: 'anā]
 [V: nāwala, N: huwa]

VV : The VV pair realizes at sentence as well as at phrasal level an inversed Head-Modifier relation.¹² We may distinguish, at least, a modal (example 15), a temporal (example 16) and a negative relation (example 17).

- (15) ġa'ala yuḥaddiṭu nafsahu
 began-he he-talks self-his
 He began talking to himself (Cant.:II,163,7)
 [V: ġa'ala, N: huwa]
 [V: ġa'ala, V: ḥaddaṭa] [V: ḥaddaṭa, N: huwa]
 [V: ḥaddaṭa, N: nafs] [N: nafs, N: huwa]
- (16) kāna qad ġāwaza 'lsittina
 was-he certainly passed-he the-sixty
 He had passed over sixty (Cant.:I,72,2)
 [V: kāna, N: huwa] [V: kāna, V: ġāwaza]
 [V: ġāwaza, N: huwa]
 [V: ġāwaza, N: sittina]

⁸ The source indication refers to: El-Ayoubi et al, Teil w, Band x, page y, example z.

⁹ Here and in the following example the personal pronoun is mapped into its independent form.

¹⁰ The modifier is realized in the form of an asyndetical relative clause.

¹¹ In this example we have a suffixed pronoun as benefactive object together with a direct object preceded by the object-introducer 'iyyā.

¹² The 'English' approach disregards so-called 'auxiliaries' where the 'Arabic' approach speaks about 'non-coinciding phrasal heads'.

- (17) lastu ʿuḥibbuhum
 was-not-I I-love-them
 I do not love them (Cant.:I,124,8)
 [V: laysa, N: ʿanā] [V: laysa, V: ʿaḥabba]
 [V: ʿaḥabba, N: ʿanā] [V: ʿaḥabba, N: huwa]

This analysis shows that the HM-relations in 5 The IR-tools for Arabic text data in Arabic differ strongly from those in English (compare the figures 1 and 6).

5 The IR-tools for ArabicText Data

So far we discussed the rationale behind information retrieval from English and Arabic text data within an almost identical linguistic and a fully identical computational framework. The IR-tools for English are available and the system is operational. What about the Arabic IR-tools? As far as we know, a fully operational academic or commercial Arabic IR-system consisting of a parser, a lexicon and a corpus is not (yet) available. So we have to rely on ourselves.

As far as the parser is concerned, a parser used for the automatic syntactic parsing of the noun phrase and the verb phrase in Modern Standard Arabic (Ditters, 1992) has been extended to cover sentence level. This parser is currently being updated towards the requirements of a new version of the AGFL-formalism. At the same time a subset (ARP4IR) of this parser is being developed for the IR of Arabic text data.

In the framework of the DIINAR-MBC project, supported by the European Union (INCO-DC N° 961 791), tools for the processing of Arabic text data have been developed. We will use the different lexicons, compiled within this project, for our IR-purposes.

An ideal test corpus is the 1991 year edition of categorized newspaper articles of al-Hayat. The data have been distributed into subject-specific databases, thus following the Al-Hayat subject tags: General, Car, Computer, News, Economics, Science, and Sport. Mark-up, numbers, special characters and punctuation have been removed. We will use this sub-corpus for test purposes. At a later stage we intend to use other parts of the al-Hayat corpus, the size of which surpasses the 268 MB and contains 18,639,264 distinct tokens in 42,591 articles, organized in 7 domains.

6 Conclusion

We have shown how the aboutness of Arabic phrases (and clauses) may be expressed in the form of a transduction to Head/Modifier Pairs, including lemmatization, analogous to the transduction schema for English phrases (Koster, 2004). The transduction described can be integrated without problems into the formal grammar for MSA developed by Ditters, using the transduction mechanism of the AGFL formalism.

Although the same set of significant pairs (NN, NV, VN and NA, as well as A, V and N with Preposition Phrases as modifiers) can be extracted from both Arabic and English

text, there are conspicuous differences in the relations embodied by those pairs.

As an example, in Arabic both the subject- and the object-relation may be realized by a VN pair, so that the polarity between them disappears. In principle, all such conflicts can be resolved by marking the pairs for the relation they represent. It is also possible to express the object relation by a VN pair and the subject relation by an NV pair, by interchanging the elements as in English, but this would lead to confusion with the topic-comment relation unless the inverted order has formally been marked.

Furthermore, due to the strong orientation of Modern Standard Arabic to predicative sentences and participial constructions, for representing Arabic text also VV pairs might appear to be necessary, in distinction to English.

It is clear that much work is still ahead of us before a set of language-independent syntacto-semantical relations has been found, expressed in a uniform way as (possibly nested) HM pairs. But even without such an ideal framework, the HM pairs described here may be used in many applications to enrich the traditional bag-of-words representation of documents in Arabic Information Retrieval.

Our next step will be the determination of the relative frequencies of those pairs in large Arabic corpora, and the relative contribution of these various relations to the aboutness of documents. We will evaluate the possible variants of this document representation by means of a large scale experiment in the automatic classification of Arabic newspaper articles from al-Hayat.

We expect that this new document representation will find many practical applications in Arabic Information Retrieval.

7 References

- Arampatzis, A., T. Tsores, C.H.A. Koster and Th.P. van der Weide. 1998. 'Phrase-based Information Retrieval' in: *Information Processing & Management Journal*, Vol. 3, 34, no 6, December 1998, pp. 693-707.
- Arampatzis, A., Th.P. van der Weide, C.H.A. Koster, and P. van Bommel. 2000. 'An Evaluation of Linguistically-motivated Indexing Schemes' in: *Proceedings BCS-IRSG 2000 Colloquium on IR Research*, Cambridge, England.
- Bolshakov, Igor A. 2004. 'Getting One's First Million ... Collocations' in: Gelbukh, A. (Ed.) *Computational Linguistics and Intelligent Text Processing*, (CICLing-2004). Springer LNCS 2945 p. 229-242.
- Bruza, P, and T.W.C. Huibers. 1996. 'A Study of Aboutness in Information Retrieval' in: *Artificial Intelligence Review*, 10, p 1-27.
- Cantarino, Vicente. 1974-5. *Syntax of Modern Arabic Prose*. 3 Vols. Bloomington: Indiana University Press.
- Ditters, E. 1991. 'A Modern Standard Arabic Sentence Grammar' in: *Bulletin d'Études Orientales*, Paris, 43, pp. 197-236.

- Ditters, E. 1992. *A Formal Approach to Arabic: the Noun Phrase and the Verb Phrase*, PhD thesis, Humanities Department Nijmegen University, Nijmegen: Luxor.
- Ditters, E. 2001. 'The Description of Modern Standard Arabic Syntax in terms of Functions and Categories' in: *Langues et Littératures du Monde Arabe*, 2 (2001), pp.115-151.
- Ditters, E. 2003a. 'Non-coinciding Phrasal Heads' in *Proceedings of the Joint International Conference on Computer, Communication and Control Technologies (CCCT'03) and the 9th International Conference on Information Systems Analysis and Synthesis (ISAS'03)*, July 31-August 2, Orlando Florida, USA, VOL. IV, pp. 51-56.
- Ditters, E. 2003b. 'An AGFL for the Description of Non-coinciding Phrasal Heads' in *Proceedings of the Joint International Conference of the 7th World Multiconference on Systemics, Cybernetics and Informatics (SCI 2003)*, July 28-30, Orlando Florida, USA, VOL. VI, pp. 107-112.
- El-Ayoubi, H., W. Fischer and M. Langer. 2001. *Syntax der Arabischen Schriftsprache der Gegenwart*, Teil 1, Band 1, Reichert Verlag: Wiesbaden.
- El-Ayoubi, H., W. Fischer and M. Langer. 2003. *Syntax der Arabischen Schriftsprache der Gegenwart*, Teil 1, Band 2, Reichert Verlag: Wiesbaden.
- Evans, D.A., R.G. Lefferts, G. Grefenstette, S.H. Handerson, W.R. Hersch and A.A. Archbold. 1993. 'CLARIT TREC Design, Experiments and Results' in: *TREC-1 proceedings*, pp. 251-286.
- Fagan, J.L. 1988. *Experiments in automatic phrase indexing for document retrieval: a comparison of syntactic and non-syntactic methods*, PhD Thesis, Cornell University.
- Gelbukh, A., G. Sidorov, S.-Y. Han, and E. Hernández-Rubio. 2004. 'Automatic Syntactic Analysis for Detection of Word Combinations' in: Gelbukh, A. (Ed.): *Computational Linguistics and Intelligent Text Processing (CICLing-2004)*. Springer LNCS 2945, Heidelberg, 2004, pp.
- Koster, C.H.A. 1991. 'Affix Grammars for Natural Languages' in: Alblas, H., and B. Melichar (Eds.): *Attribute Grammars, Applications and Systems*. Springer LNCS 545, Heidelberg, pp. 469-484.
- Koster, C.H.A. and M. Seutter. 2002. 'Taming Wild Phrases' in: *Proceedings 25th European Conference on IR Research (ECIR 2003)*, Springer LNCS 2633, Heidelberg, pp 161-176.
- Koster, C.H.A. 2004. 'Head/Modifier Frames for Information Retrieval' in: *Proceedings CICLing-2004*, Springer LNCS 2945, Heidelberg, pp 420-432.
- Lewis, D.D. 1992. *Representation and Learning in Information Retrieval*. PhD thesis, Department of Computer Science, Univ. of Massachusetts, Amherst, MA 01003.
- Sparck Jones, K. 1999. 'The role of NLP in Text Retrieval' in: Strzalkowski, T. 1999. pp. 1-24.
- Smeaton, A.F. 1997. 'Using NLP and NLP resources for Information Retrieval Tasks' in: Strzalkowski, T. (Ed.): *Natural Language Information Retrieval*, Kluwer Academic Publishers, ISBN 0-7923-5685-3, pp.
- Strzalkowski, T. 1995. 'Natural Language Information Retrieval' in: *Information Processing and Management*, 31 (3), pp. 397-417.
- Strzalkowski, T. (Ed). 1999. *Natural Language Information Retrieval*, Kluwer Academic Publishers, ISBN 0-7923-5685-3.