==
Language in India www.languageinindia.com ISSN 1930-2940 Vol. 18:8 August 2018
India's Higher Education Authority UGC Approved List of Journals Serial Number 49042
==

Grammatical Relations in Arabic Compound Words: Evidence from Corpus-linguistics

Mohammed Modhaffer and Dr. C. V. Sivaramakrishna

==

Abstract

This paper investigates the grammatical relations in Arabic bigram compound words in the frame work of Scalise and Bisetto (2009). Total of 16570 compound words were extracted from more than 672 million words, using contingency tables and log-likelihood ratio. Data analysis revealed that the ranking order of the grammatical relations is as follows: attributive (51.79%), subordination (47.70%) and coordination (0.51%).

1. Introduction

Arabic is the Semitic language spoken by circa 400 million native speakers in the Middle East and it is also the formal language in the religious functions of more than one billion Muslims around the world. Arabic is also one of the 6 languages of the United Nations. Standard Arabic is composed of Classical Arabic (CA) and Modern Standard Arabic (MSA). CA is the variety of the Holy Qur'an. It served as the medium of communication, literature, trade and commerce during the golden era of Islamic Empire (7th Century – 13th Century circa). MSA is a revival copy of CA and it came into existence in the 19th Century. In terms of spelling and morphology, MSA does resemble CA to a large extent, but both differ in terms of structure, where MSA is said to use a simpler structure. For instance, the following structure longer appears in MSA texts:

(1) ʔaʕtˤaj-ta-ni:-ha:
give.PAST-2.SG.MASC-NOM-me.ACC-it.ACC
"You gave it to me."

Instead, in MSA the above structure is expressed in a way similar to the following:

(2)	ʔanta	ʔaʕtˤaj-ta-ha:	l-i:
	You.2.SG.MASC.NOM	give.PAST-2.SG.MASC-NOM-it.FEM-ACC	to-me.GEN
		"You gave it to me".	

In the present-day educational system of Arab world, MSA is learnt at elementary and upper-elementary school onwards, while CA is learnt at higher levels of education such as graduate

The abbreviations used throughout this paper are as follows: MSA = Modern Standard Arabic, CA = Classical Arabic, SG = singular, MASC = masculine, FEM = feminine, DEF = definite article, NOM = nominative case, ACC = accusative case, GEN = genitive case, NLP = Natural Language Processing, 1 = first person, 2 = second person, 3 = third person, PAST = past simple tense, ∅ = zero case assignment.

and post-graduate programs in Arabic language and literature. MSA is learnt explicitly through textbooks and it is rather difficult. However, CA is learnt explicitly through classical books and manuscripts which date back to the 7th Century (i.e. Seebawayh's era) and it is very hard to learn even for native speakers of Arabic. A substantial part of the vocabulary of CA, which had been employed by the Abbasid author Al-Jaħiẓ (died 869), is no longer employed by any contemporary Arab author. A native speaker of Arabic pursuing a post-graduate program in Arabic literature would hardly understand the books of Al-Jaħiẓ without recourse to **Lisaanul Arabi** – the standard dictionary of Arabic.

Whatever be the case, an overlap exists between CA and MSA in terms of lexicon and structure. This may be attributed to the fact that the Holy Qur'an is still read and learnt by every (Muslim) native speaker of Arabic. That is, the Holy Qur'an and the huge body of religious and literary texts which are written in CA have served as an archive for CA. For more information on both CA and MSA, see Versteegh (2014), Watson (2002), Bateson (1967) and Al-Huri (2015).

Morphologically speaking, Arabic is highly inflectional with a root and pattern morphology and much overlapping of morphological features. Syntactically speaking, Arabic is a pro-drop language with two different word orders, of which the unmarked is Verb + Subject + Object. Having such inflectional morphological status and a pro-drop syntactic nature, Arabic poses severe challenges to natural language processing (NLP) in all levels of linguistic analysis.

The remainder of this paper is organized as follows: Section (2) briefly surveys the related literature. A description of the text corpora in terms of counts and genres, method of extraction and filtering is presented in Section (3). Section (4) presents data analysis and cites examples of the grammatical relations in Arabic compound words. Section (5) concludes this paper with conclusions.

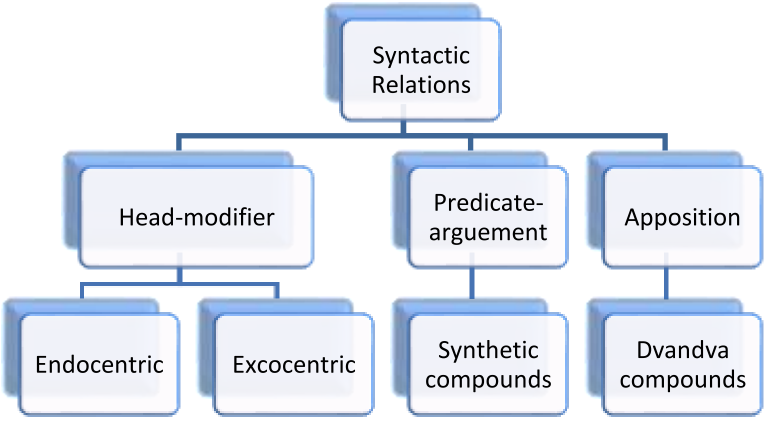
2. Brief Literature Background

Several definitions of compounds have been proposed by different authors. The simplest one is that of Fabb (2001) who defines a compound as "a word which consists of two or more words" (p. 66). In a similar fashion, Montermini (2010: 30) states that "it is commonly admitted that a prototypical instance of compounding is the product of the combination of more than one word".

Scalise & Vogel (2010: 5) list different definitions of compounds based on their basic building blocks. For instance, considering root as the basic building blocks, compounds are best defined as "combinations of two or more roots" (cf. Harley (2009) and Katamaba (1993)). For others, the basic building blocks are lexemes. According to this view, compounds may be defined as a combination of two or more lexemes, each of which can function as an independent lexeme (cf. Bauer (2001), Haspelmath (2002) and Booij (2005)).

Spencer (1991: 310) states that "… the elements of a compound may have relations to each other which resemble the relations holding between the constituents of a sentence. The three important relations are head-modifier, predicate-argument and apposition". The following figure summarizes these relations:

Figure 1: Syntactic relations between the constituents of the compounds



3. Methodology

Total of 16570 compound words were automatically extracted from Classical Arabic and Modern Standard Arabic multi-genre text corpora of 672,242,076 words, of which 473,498,083 tokens are Classical Arabic texts and the remaining 198,743,993 words are Modern Standard Arabic texts. Table (1) shows the genres and their counts.

Table 1: corpora genres and counts

S.N.	Genre	Number of words	Variety
1	History	40,272,729	CA
2	Holy Qur'an Explanation	102,517,668	CA
3	Jurisprudence	114,723,632	CA
4	Literature	38,128,323	CA
5	Prayers	45,165,305	CA
6	Prophet's Biography	24,481,634	CA
7	Prophet's Sayings	86,714,442	CA
8	Standard Arabic Lexicons	21,494,350	CA
9	Defense	21,020,880	MSA
10	Encyclopedic texts	13,254,157	MSA
11	Information technology	11,650,339	MSA
12	Law	15,242,340	MSA
13	Medical Texts	13,684,449	MSA
14	Miscellaneous Science Texts	6,380,333	MSA
15	Newswire	117,511,495	MSA
	Grand total of CA and MSA	**672,242,076**	

CA texts were extracted from 5000 e-books belonging to the Shamela Library which can be obtained for free[2]. The Shamela Library was classified by human classifiers. The e-books were converted into UTF-8 text files. Then the texts were cleaned from punctuation marks, vocalization marks (diacritics) and symbols.

MSA newswire texts were retrieved from the corpus collected by Dr. Ahmed Abdelali. It contains 113 million tokens and it can be obtained for free[3]. The remaining four million tokens of newswire as well as the remaining genres were crawled from the World Wilde Web. By default, MSA texts are not vocalized, and the punctuation marks as well as symbols were simply stripped at the time of crawling.

Before extracting the candidate constructions, we trained our own model of Stanford Part of Speech Tagger (Toutanova and Manning , 2000; Toutanova, Klein, Manning, & Singer, 2003) and tagged the above-mentioned corpora. It has to be noted that the overall accuracy of the model is 95.52% and 81.45% on unknown words. For the sake of morphological analysis, we used our own rule-based morphological analyzer to separate prefixes and affixes from Arabic words in the text corpora.

Quantification of Arabic compound words was conducted using contingency tables and log-likelihood ratio as in the work of Seretan (2011). Following the extraction, we manually filtered out the false positive compounds. Then the final true compounds were exported into a Structured Query Language (SQL) database. In the database, compound words were analyzed manually, and the grammatical relations were worked out.

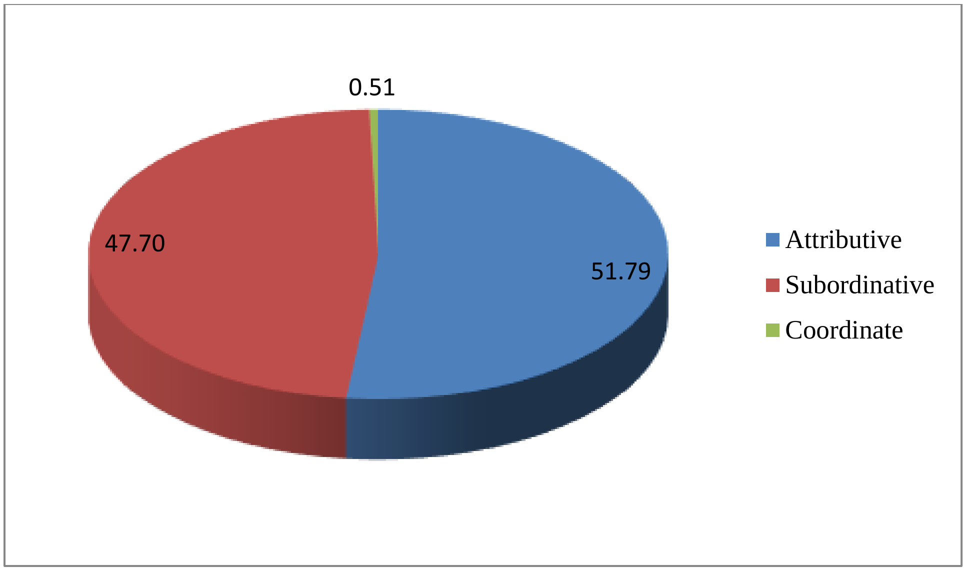
4. Data Analysis

Following the approach of Scalise and Bisetto (2009), there are three grammatical relations holding between the constituents of the compounds in our database. Figure (2) plots the distribution.

Figure 2: Distribution of grammatical relations between the constituents of compound words following Scalise and Bisetto (2009)

[2] http://sourceforge.net/projects/albahhet/files/ShamelaEpub/shamela_epub.tar.gz/download

[3] http://aracorpus.e3rab.com/argistestsrv.nmsu.edu/AraCorpus.tar.gz



Attribution grammatical relation holds between the constituents of more than half of the compounds in our database (8581 out of 16570 compound words). The following examples illustrate attribution grammatical relation in Arabic compounds:

(3) ʔiʕtima:d-un mustanadijj-un
credit.SG.MASC.INDEF-NOM of document.SG.MASC.INDEF-NOM
'letter of credit'

(4) xuma:sijj-u t-taka:fuʔ-i
of five.SG.MASC.INDEF-NOM DEF-valency.SG.MASC-GEN
'pentavalent'

(5) ʔab-un ru:ħijj-un
father.SG.MASC.INDEF-NOM spiritual.SG.MASC.INDEF-NOM
'spiritual leader'

(6) ʔiba:dat-un ʒama:ʕijjat-un
killing.SG.FEM.INDEF-NOM mass.SG.FEM.INDEF-NOM
'genocide'

(7) ʔaʒr-un ʔismijj-un
wage.SG.MASC.INDEF-NOM nominal.SG.MASC.INDEF-NOM
'nominal wage'

Example (3) is a noun modified by an adjective and it is common in banking and finance texts. Example (4) is an adjective in nature _ it is composed of an adjective in the X slot and a definite noun in the Y slot, and it is common in Chemistry texts. Examples (5 – 7) are all nouns modified by adjectives. Example (5) is common in religious or social texts. Example (6) is common in legal, political and newswire texts. Example (7) is common in administrative and legal texts.

Total of 7904 compound words (circa 47.70%) exhibited subordination grammatical relation between their constituents. For instance,

(8) waki:l-u l-waza:r-at-i
agent.SG.MASC.INDEF-NOM DEF-ministry.MASC-FEM-GEN
'undersecretary'

(9) mawɗiʕ-u ʃ-ʃakk-in
place.SG.MASC.INDEF-NOM doubt.SG.MASC.INDEF-NOM
'questionable'

(10) ʔistiʕa:dat-u n-niʑa:m-i
restoring.SG.FEM.INDEF-NOM DEF-system.SG-MASC-GEN
'system restore'
(11) tahri:b-u l-baʃar-i
smuggling.SG.FEM.INDEF-NOM DEF-human.MASS-MASC-GEN
'human being smuggling'

(12) baħθ-u ʕamalijja:t-in
research.SG.MASC.INDEF-NOM operation.PL.FEM.INDEF-GEN
'operation research'

According to CA grammatical theory, the first constituents in Examples (8 – 12) are called muɗa:fun 'added' and the second constituents are called muɗs:fun ʔilajhi 'the destination to which the first constituent is added to'. In modern linguistic theory, however, things are the other way around. That is, the second constituents are subordinate to the first constituents. That is, the first constituent is the head of the compound word, and such it is dominates and governs whatever constituents come under it.

The last and least grammatical relation attested in the compounds in our database is coordination. It was present in only 85 compounds (circa 0.51%). For example,
(13) ʔiliktrun fult
electron.SG.MASC.INDEF-ø volt.SG.MASC.INDEF-ø
'electron volt'

(14) hajdruksi:d-u sˤu:dju:m
hydroxide.SG.MASC.INDEF-ø sodium.SG.MASC.INDEF-ø
'sodium hydroxide'

Example (13) is composed of electron and volt and both of these words equally contribute to the total meaning of the compounds. Similarly, Example (14) is composed of hydroxide and sodium and both words contribute equally to the total meaning of the whole compound. It has to be noted that Example (13) is neither electron nor volt and Example (14) is neither hydroxide nor sodium.

5. Conclusions

Following the approach of Scalise and Bisetto (2009), grammatical relations were worked out in our database: **attributive**, **subordinate** and **coordinate**. Attributive grammatical relation was found present in 51.79% (8581 out of 16570). Attributive grammatical relation is by far the most frequent grammatical relation in our database. This can be straightforwardly explained by the fact that 8024 compounds (circa 48.42%) had an adjective in one of their constituents.

The second top grammatical relation attested in our database is subordination. In this relation, the non-head is subordinate to the head of the compound. It has to be made clear that in CA grammatical theory, the head is subordinate to the non-head. That is because the head is considered *muɗa:fun* and the non-head is *muɗa:fun ʔilajhi*. Subordinate grammatical relation scored 47.70%.

Coordinative grammatical relation was the least one to be attested in our database, with only 0.51%. In these compounds, neither X modifies Y or the vice versa, and neither X is subordinate to Y or the vice versa, and both X and Y constituents equally contribute the total meaning of the whole compounds. This conforms to the results of the survey conducted by Wälchli (2005: 215) who placed Arabic in the lowest level in terms of presence of compound words in the languages of Europe and Asia. Arabic is placed in the zero level which means that co-compounds almost do not exist. It has to be noted that coordination can be used as grammatical and semantic criterion for classifying compound words.

===

References

Al-Huri, I. (2015). Arabic Language: Historic and Sociolinguistic Characteristics. *English Language and literature Review, 1*(4).

Bateson, M. (1967). *Arabic Language Handbook.* Georgetown University Press.

Bauer, L. (2001). Compounding. In M. Haspelmath, E. Konig, W. Oesterreicher, & W. Raible (Eds.), *Language Typology and Language Universals: An International Handbook* (pp. 695-707). Berlin: Mouton de Gruyter.

Booij, G. (2005). Compounding and Derivation: Evidence for construction morphology. In W. Dressler, D. Kastovsky, O. Pfeiffer, & F. Rainer (Eds.), *Morphology and Its Demarcations: Selected Papers from the 11th Morphology Meeting* (pp. 109–132). Amsterdam: John Benjamins Publishing Company.

Fabb, N. (2001). Compounding. In A. Spencer, & A. Zwicky (Eds.), *The Handbook of Morphology* (pp. 66-83). Blackwell.

Harley, H. (2009). Compounding in Distributed Morphology. In R. Lieber, & P. Štekauer (Eds.), *The Oxford Hanbook of Compounding* (pp. 129-144). Oxford University Press.

Haspelmath, M. (2002). *Understanding Morphology*. London: Arnold.

Katamaba, F. (1993). *Morphology*. London: Macmillan.

Montermini, F. (2010). Units in Compounding. In S. Scalise, & I. Vogel (Eds.), *Cross-Disciplinary Issues in Compounding* (pp. 77-92). Amsterdam: John Benjamins Publishing Company.

Scalise, S., & Bisetto, A. (2009). The Classification of Compounds. In R. Lieber, & P. Stekauer (Eds.), *The Oxford Handbook of Compounding* (pp. 49 - 80). Oxford University Press.

Scalise, S., & Vogel, I. (2010). *Cross-Disciplinary Issues in Compounding*. Amsterdam: John Benjamins Publishing Company.

Seretan, V. (2011). *Syntax-based Collecation Extraction*. Springer.

Spencer, A. (1991). *Morphological Theory: An Introduction to Word Structure in Generative Grammar*. Wiley.

Toutanova, K. and C. Manning. (2000). Enriching the Knowledge Sources Used in a Maximum Entropy Part-of-Speech Tagger. *In Proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora (EMNLP/VLC-2000)*, (pp. 63-70).

Toutanova, K., Klein, D., Manning, C., & Singer, Y. (2003). Feature-Rich Part-of-Speech Tagging with a Cyclic Dependency Network., (pp. 252-259).

Versteegh, K. (2014). *The Arabic Language*. Edinburgh University Press.

Wälchli, B. (2005). *Co-compounds and Natural Coordination*. Oxford University Press.

Watson, J. (2002). *he Phonology and Morphology of Arabic*. Oxford University Press.

===

MOHAMMED MODHAFFER (Corresponding author)
Ph.D. Research Scholar
Department of Linguistics
Kuvempu Institute of Kannada Studies
University of Mysore
Manasagangotri
Mysore – 570006
Karnataka
India
modhaffer@gmail.com

===

ORCID iD: http://orcid.org/0000-0001-7866-418X
QR Code:

DR. C.V. SIVARAMAKRISHNA (Co-author)
Research Guide
Reader-cum-Research Officer
Central Institute of Indian Languages
Ministry of Human Resource Development, Government of India
Hunsur Road, Manasagangotri
Mysore – 570006
Karnataka
India
shivaramakrishna1963@gmail.com