==
Language in India www.languageinindia.com ISSN 1930-2940 Vol. 17:9 September 2017
UGC Approved List of Journals Serial Number 49042
==

Distribution of Simple Prepositions in Modern Standard Arabic

Mohammed Modhaffer and C.V. Sivaramakrishna

==

Abstract

This paper aims to reveal the frequency distribution of simple prepositions in Modern Standard Arabic (MSA or Arabic, for short). We investigate multi-genre text corpus of 106,572,775 words. We tag the corpora with our own trained model of Stanford Part of Speech Tagger and we use our own morphological analyzer to separate the prefixes and suffixes from the tagged corpora. Results reveal that 55 prepositions constitute 16.7987% of the total vocabulary of Modern Arabic texts. Every sixth word in Arabic is a preposition. Moreover, the five most commonly used prepositions in Arabic are /li/ 'for', /fi:/ 'in', /bi/ 'with, by' /min/ 'from', and /ʕalaa/ 'on'; together, they represent 76.5550% of all the occurrences of prepositions and they cover 12.8603% of the total words in the whole corpus.

Keywords: prepositions, distribution, text corpus, Modern Standard Arabic, Semitic languages

1. Introduction

Prepositions constitute one of the core grammatical categories of Arabic vocabulary. Every sixth word in Arabic texts is a preposition. Prepositions are used to indicate several functions such as location, time, relation, instrumentation, cause and effect and so on. For a complete list of the meanings and functions of Arabic prepositions in Classical Arabic (CA) grammatical tradition, see Al Shumasan (1987).

Most of Arabic prepositions are unigram words while some of them are in the form of affixes which can be prefixed into all types of nouns. The aim of this paper is to examine the frequency distribution of simple prepositions in Modern Standard Arabic (MSA) text corpora. By "simple" we mean the prepositions which are either prefixes such as /li/ 'for' or those which are composed of only one word such as /min/ ‘from’. Complex prepositions such as /*bir-raʁmi min/* ‘despite’ fall outside the scope of this paper.

2. Brief Literature Review

English prepositions received a considerable amount of research using corpus-based studies. Roslim and Mukundan (2011) presented a good overview of the corpus-based studies on prepositions in British and American English. They stated that early works were based on the Brown corpus and Lancaster-Oslo/Bergen (LOB) corpus.

Mindt and Weber (1989) compared the frequency distribution of prepositions in British English and American English. They used the Brown corpus for American English and LOB corpus for British English. Each one of the corpora consisted of 1 million word. The authors concluded that there was no significant difference in the frequency distribution of prepositions in British English and American English. Further, they found that the most common six prepositions in American and British English are the same: of, in, to, for, with, on.

The case is rather different in Arabic. Prepositions in Arabic received little attention generally and only a few papers investigated them using corpus-based approach. Alotaiby, et al. (2014) compared the distribution of diacritics distribution, word-length, paragraph length and n-grams in Arabic and English. They showed that Arabic exceeded English in a number of parameters such as word types and bigram tokens. However, the authors reported that Arabic text corpora suffered several shortcomings such as poor organization and spelling errors.

Green (2009) investigated improving parsing performance for Arabic prepositional phrase attachment ambiguity. His best feature set achieves 80.14% F1, a 1.47% improvement over the baseline. He could gain a 7.7% F1 improvement in Arabic construct noun phrase attachment contexts.

Shilon, et al. (2012) investigated incorporating linguistic knowledge in statistical machine translation to translate prepositions from Arabic to Hebrew. They used monolingual language resources to determine the set of prepositions that are most likely to occur with each verb. They found that incorporating such knowledge significantly improved the translation of prepositions from Arabic to Hebrew.

Saeed (2014) examined the syntax and semantics of Arabic spatial prepositions. She argued that Arabic prepositional elements can be divided into the two main spatial domains: place and path. She showed that most of the path elements are mono-morphemic.

Most of the above-cited studies on Arabic prepositions focused on narrow and specific aspects of Arabic prepositions such as machine translation, parsing and disambiguation, while the remaining studies followed purely theoretical approaches in their investigation.

Despite being simple in nature, a corpus-based study of the frequency distribution of prepositions in Arabic has not yet been conducted. The reasons might be attributed to the fact that Arabic prepositions are polysemous in nature. Some prepositions can function as adverbs too. Further, some prepositions are homographs with words belonging to other closed-class categories such as relative pronouns. For example, “من” possesses two readings: /min/ “from” and /man/ “who”. The situation is further complicated by the fact that some of Arabic prepositions are prefixes to nouns and the separation of which is a very difficult task for the state of the art Arabic morphological analyzers. The top reasons for not studying frequency distribution of Arabic prepositions include: 1) the absence of a reliable part of speech tagger’s model which is capable of learning the different readings of homograph words and 2) the untackled challenges of separating the prepositions which are written as prefixes.

3. Methodology

The prepositions were automatically extracted from MSA multi-genre text corpora. Table 1 gives details of the corpora genres and counts.

Table 1: Details of the text corpora

S.N.	Genre	Count
1.	Arabic Encyclopedia	12,074,459
2.	Information Technology	10,642,705
3.	Law	13,990,679
4.	Medicine	12,550,449
5.	Military	18,984,193
6.	Newswire	38,330,290
	Total	**106,572,775**

We trained our own model of Stanford Part of Speech Tagger (2000) and tagged the above mentioned corpora. It has to be noted that the accuracy we achieved is 95.591010%. 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.

Following the Penn Treebank Tagset, all prepositions listed in table 2 were tagged as "IN". The extraction process is simple and straightforward. A loop was used to extract all the words bearing the tag "IN" and count their frequency. Finally, the frequency of occurrence and relative frequency of each preposition were calculated.

4. Data Analysis

This section provides details on the distribution of Arabic prepositions as attested in the actual usage of the text corpora. Table 2 lists all the prepositions, frequency of occurrence and relative frequency in descending order from most frequent to least frequent. The transcription of Arabic grapheme strictly follows the guidelines of the International Phonetic Alphabet (IPA). "Frequency of occurrence" means how many times a given preposition was observed in the text corpora. "Relative frequency" gives the percentage of each preposition relative to the total occurrences of all prepositions. "Frequency % relative to the total corpus" gives the percentage covered by a given preposition relative to the total amount of words in the corpus which is **106,572,775** words.

Table 2: Rank List of Prepositions in Modern Standard Arabic

S.No.	Arabic Script	Transcription and Gloss	Frequency of occurrence	Relative frequency %	Frequency % relative to the total corpus
1	ل	/li/ ‘for’	3233469	18.0612	3.0340
2	في	/fi:/ ‘in’	3116311	17.4068	2.9241
3	ب	/bi/ ‘with, by’	2958555	16.5256	2.7761
4	من	/min/ ‘from’	2616431	14.6146	2.4551
5	على	/ʕalaa/ ‘on’	1780773	9.9469	1.6709
6	إلى	/ʔilaa/ ‘to’	934553	5.2201	0.8769
7	عن	/ʕan/ ‘about’	667658	3.7293	0.6265
8	مع	/maʕa/ 'with'	413222	2.3081	0.3877
9	ك	/ka/ ‘as’	326440	1.8234	0.3063
10	بين	/bajn/ 'between'	254109	1.4194	0.2384
11	بعد	/baʕd/ 'after'	219442	1.2257	0.2059
12	خلال	/xilaal/ 'through, during'	154573	0.8634	0.1450
13	قبل	/qabl/ 'before'	152257	0.8505	0.1429
14	حتى	/ħattaa/ ‘till’	140701	0.7859	0.1320
15	عند	/ʕind/ 'at'	97190	0.5429	0.0912
16	حول	/ħawl/ 'around'	93652	0.5231	0.0879
17	تحت	/taħt/ 'under'	91417	0.5106	0.0858
18	لدى	/ladaa/ 'at'	66334	0.3705	0.0622
19	منذ	/munð/ ‘since’	66071	0.3691	0.0620
20	عبر	/ʕabr/ 'through'	45790	0.2558	0.0430
21	أمام	/ʔamaam/ 'in front of'	41297	0.2307	0.0388
22	نحو	/naħw/ 'around, towards'	38378	0.2144	0.0360
23	و	/wa/ ‘and’ (in swearing)	36483	0.2038	0.0342
24	داخل	/daaxil/ 'in, inside'	35648	0.1991	0.0334
25	وفق	/wifq/ 'according to, as'	34680	0.1937	0.0325
26	حسب	/ħasb/ 'as, according to'	34478	0.1926	0.0324
27	ضد	/ɗid/ 'against'	29335	0.1639	0.0275

28	ضمن	/ɗimn/ 'in, including'	27451	0.1533	0.0258
29	فوق	/fawq/ 'on'	19847	0.1109	0.0186
30	مقابل	/muqaabil/ 'opposite to'	18974	0.1060	0.0178
31	حوالي	/ħawaalaj/ 'around'	18170	0.1015	0.0170
32	خارج	/xaariɟ/ 'out'	17255	0.0964	0.0162
33	وسط	/wast̥/ 'in middle of'	17009	0.0950	0.0160
34	دون	/duun/ 'without'	14871	0.0831	0.0140
35	بحسب	/bi-ħasb/ 'according to'	14402	0.0804	0.0135
36	أعلى	/ʔaʕlaa/ 'up'	11350	0.0634	0.0106
37	قرب	/qurb/ 'near'	10504	0.0587	0.0099
38	إثر	/ʔiθr/ 'after, because'	9937	0.0555	0.0093
39	كي	/kaj/ ‘as’	7463	0.0417	0.0070
40	خلف	/xalf/ 'behind'	6312	0.0353	0.0059
41	متى	/mataa/ ‘from’	6075	0.0339	0.0057
42	عقب	/ʕaqib/ 'after'	6038	0.0337	0.0057
43	وراء	/waraaʔ/ 'behind'	5672	0.0317	0.0053
44	عدا	/ʕadaa/ ‘except’	4219	0.0236	0.0040
45	إبان	/ʔibaan/ 'during'	2037	0.0114	0.0019
46	لعل	/laʕalla/ ‘may’	1876	0.0105	0.0018
47	أسفل	/ʔasfal/ 'down'	1092	0.0061	0.0010
48	تلو	/tilw/ 'after'	1053	0.0059	0.0010
49	بدل	/badal/ 'instead of'	717	0.0040	0.0007
50	خلا	/xalaa/ ‘except’	663	0.0037	0.0006
51	رب	/rubba/ ‘may’	308	0.0017	0.0003
52	مذ	/muð/ ‘since’	150	0.0008	0.0001
53	لقاء	/liqaaʔ/ 'in return of'	70	0.0004	0.0001
54	حاشا	/ħaaʃaa/ ‘except’	59	0.0003	0.0001
55	ت	/ta/ ‘by’ (in swearing)	45	0.0003	0.00004
		Total	**17902866**	**100%**	**16.7987%**

Table 2 shows that the most frequent preposition in Arabic is /li/ ‘for’ and the least frequent preposition is /ta/ ‘by (in swearing)’. In order to dive deeper into the distribution of prepositions in Arabic, table 3 shows the central and marginal prepositions which a divided into seven groups.

Table 3: Groups of Prepositions in Modern Standard Arabic

Group	Range of frequency	Number of prepositions	Absolute frequency	Relative frequency %	Cumulative frequency %
Group 1	Over 3,000,000	2	6349780	35.4680	35.4680
Group 2	2,000,001 – 3,000,000	2	5574986	31.1402	66.6081
Group 3	1,000,001 – 2,000,000	1	1780773	9.9469	76.5550
Group 4	100,001 – 1,000,000	9	3262955	18.2259	94.7809
Group 5	10,001 – 100,000	23	880586	4.9187	99.6996
Group 6	1001 – 10,000	11	51774	0.2892	99.9888
Group 7	1 – 1000	7	2012	0.0112	**100**
	Total	**55**	**17902866**	**100%**	

Table 3 shows that the central prepositions in Arabic are those which fall in the first three groups. These central prepositions are five in number: /li/ ‘for’, /fi:/ ‘in’, /bi/ ‘with, by’ /min/ ‘from’, and /ʕalaa/ ‘on’. Group 4 and 5 fall in-between the two extremes. They comprise 32 prepositions whose frequency of occurrence ranges between 10,000 and less than 1 million. Group 6 and 7 comprise the 18 marginal prepositions whose frequency of occurrence is less than 10,000.

5. Results and Discussions

Comparing our results with those of Mindt and Brown (1989), fifty-five prepositions were attested in our data with a frequency of 17,902,866 occurrences. Our data is 53 times bigger, and the attested frequency of Arabic propositions relative to the whole data is 73 times bigger than that of English in Mindt and Brown (1989). The number of attested prepositions is far less than that of Mindt and Brown (1989): their list contained 94 prepositions and our list contained 55. However, it has to be noted that our list is by no means exhaustive.

The most frequent preposition in Arabic is /li/ “for”. It occurs 3,233,469 times and it scores 18.0612% of the total prepositions. Moreover, it occurs 3.0340% relative to the total size of the corpus. This finding supports the manual calculation conducted by Esseesy (2010) who showed that the preposition /li/ ‘for’ occurs 3.4903 per 100,000 words. The second most frequent proposition in Arabic is /fi:/ ‘in’, with a frequency of occurrence of 3,116,311times. It constitutes 17.4068% of all the prepositions and it occurs 2.9241% relative to the total number of words in the corpus. The top two prepositions /li/ ‘for’ and /fi:/ ‘in’ situate themselves in the top group of Arabic prepositions, with over 6 million occurrences, i.e. they cover 5.9582% of all the words in the corpus. They form 35.4680% of all prepositions.

The second group of prepositions contains 2 prepositions: /bi/ ‘with, by’ and /min/ ‘from’. Together, they have a frequency of occurrence of more than 5 million occurrences and they constitute 31.1402% of all the prepositions. They constitute 5.2312% of all the words in the corpus.

The preposition /ʕalaa/ ‘on’ is exhibits a very interesting distribution as it forms the third most frequent group of prepositions. It has an occurrence of more than 1.7 million. It covers almost one tenth of all the prepositions (9.9469%). It occurs 1.6709 in a 100 words.

The five most frequent prepositions in Arabic are /li/ ‘for’, /fi:/ ‘in’, /bi/ ‘with, by’ /min/ ‘from’, and /ʕalaa/ ‘on’. They form more than three quarters of the distribution of all the prepositions (76.5550%) and each of them occurs more than a million times. These top five prepositions cover 12.8603% of all the words in our corpus.

The fourth most frequent group contains nine prepositions: /ʔilaa/ ‘to’, /ʕan/ ‘about’, /maʕaa/ ‘with’, /ka/ ‘as’, /bajn/ ‘between’, /baʕd/ ‘after’, /xilaal/ ‘through, during’, /qabl/ ‘before’ and /ħattaa/ ‘till’. This group constitutes 18.2259% of all the prepositions with a range of frequency between a hundred thousand and one million. They cover 3.0617% of the total words in the entire corpus.

The fifth group of prepositions contains 23 prepositions with a frequency of occurrence of 880586, and they form 4.9187% of all the prepositions (see Table 2 from serial number 15 to 37). These 23 prepositions cover 0.8263% of all the words in the entire corpus.

The sixth group of prepositions is the second least frequent group; comprising 11 prepositions (see Table 2 from serial number 38 to 48). These 11 prepositions have an occurrence of 51774, ranging from 1000 to 10,000. They form 0.2892% relative to all the prepositions. They cover 0.0486% of all the words in the corpus.

The seventh group contains the least frequent prepositions, and it consists of 7 prepositions: /badal/ ‘instead of’, /xalaa/ ‘except’, /rubba/ ‘may’, /muð/ ‘since’, /liqaaʔ/ ‘in return of’, /ħaaʃaa/ ‘except’ and /ta/ ‘by (in swearing)’ (see Table 2 from serial number 49 to 55). Their frequency of occurrence is 2012, ranging from 1 to 1000 times forming less than 0.02% relative to all the prepositions. They cover 0.0019% of the total words in the corpus. /rubba/ ‘may’, /muð/ ‘since’ and /ħaaʃaa/ ‘except’ are no longer active prepositions in the lexicon of Modern Standard Arabic. /muð/ ‘since’ and /ħaaʃaa/ ‘except’ are frozen to Classical Arabic texts. /badal/ ‘instead of’ and /liqaaʔ/ ‘in return of’ are currently used as nouns more than prepositions. /ta/ ‘by (in swearing)’ is still used in swearing phrases, but its use is far less than that of /wa/ ‘by (in swearing)’. /ta/ ‘by (for swearing)’ is frozen to the divine name ‘Allah’, and it is barely noticed in the MSA texts. We examined its occurrences and found most of them are from excerpts of Classical Arabic scripts or quotations from the Holy Qur’an.

6. Conclusion

In this paper, we presented the frequency of 55 simple prepositions in Modern Standard Arabic text corpora. 16.7987% of all the words in our corpus are prepositions. Prepositions in Arabic are one in six. In other words, every sixth word in Arabic is a preposition.

We ranked all prepositions from most frequent to least frequent and listed them in table 2. The most frequent preposition in Arabic is /li/ “for”. It occurs 3,233,469 times and it scores 18.0612% of the total prepositions and it occurs 3.0340% relative to the total size of the corpus. This finding supports the manual calculation made by Esseesy (2010) who showed that the preposition /li/ ‘for’ occurs 3.4903 per 100,000 words. The least frequent preposition is /ta/ ‘by (for swearing)’ which is frozen to the divine name ‘Allah’, and it is barely noticed in the MSA texts. We examined its occurrences and found most of them are from CA scripts or quotations from the Holy Qur’an.

We divided the prepositions into seven groups, based on the range of frequency of occurrence and listed them in table 3. The top three group comprise the most frequent prepositions (the central prepositions) viz. li/ 'for', /fi:/ 'in', /bi/ 'with, by' /min/ 'from', and /ʕalaa/ 'on'. The frequency of occurrence of the five most frequent prepositions makes up more than 3 quarters of the occurrences of all Arabic prepositions (76.5550%), and these five prepositions cover 12.8603%, almost one eighth, of all the words in the entire corpus.

===

References

Al Shumasan, A. (1987). *ħuruuful ɟarr: dilaalaatuhaa wa ʕilaaqaatuhaa [Prepositions: Meanings and Functions]*. Riyadh: Al-Tayyar Offset Press.

Alotaiby, F., Foda, S. and Alkharashi, I. (2014). Arabic vs. English: Comparative Statistical Study. *Arab J Sci Eng*, *39*, 809–820. doi: http://dx.doi.org/10.1007/s13369-013-0665-3

Esseesy, M. (2010). *Grammaticalization of Arabic Prepositions and Subordinators: A Corpus based Study*. Leiden: Brill.

Green, S. (2009). Improving Parsing Performance for Arabic PP Attachment Ambiguity. Retrieved August 26, 2017, from https://nlp.stanford.edu/courses/cs224n/2009/fp/30-tempremove.pdf

Mindt, D. and Weber, C. (1989). Prepositions in American and British English. *World Englishes*, 229-238. doi: http://dx.doi.org/10.1111/j.1467-971X.1989.tb00658.x

Roslim, N., and Mukundan, J. (2011). An Overview of Corpus Linguistics Studies on Prepositions. *English Language Teaching*, *4*(2), 125-131. doi: http://dx.doi.org/10.5539/elt.v4n2p125

Saeed, S. (2014). The Syntax and Semantics of Arabic PS. *Newcastle and Northumbria Working Papers in Linguistics*, *20*, 44-66.

Shilon, R., Fadida, H., and Wintner, S. (2012). Incorporating Linguistic Knowledge in Statistical Machine Translation: Translating Prepositions. *Innovative Hybrid Approaches to the Processing of Textual Data* (pp. 106-114). Avignon, France: Association for Computational Linguistics.

Toutanova, K. and Manning, C. (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.

===

Author details

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