BREADTH OF PERSIAN CORE VOCABULARY FOR URDU SPEAKERS: A PERSIAN CORPUS-BASED STUDY

***Abrar Hussain Qureshi, **Mehmood ul Hassan, ***Faiza Rehman**

ABSTRACT:

Vocabulary building of a foreign language is one of the difficult tasks for non-native learners and teachers. Traditionally, vocabulary building off a foreign language has been under the control of language teachers. Currently, there is a shift in foreign language learning from teacher's control to learner's autonomy. The idea has gained strength and the advent of computers and language corpora have facilitated this autonomy of the learners. According to Henriksen (1999), learners of a foreign language at various stages of language learning will be having different types of vocabulary in different number. The undertaken study is an attempt in this regard to facilitate Persian learners of Urdu speakers to provide them a list of core vocabulary of Persian language that will cover nearly all spheres of Persian culture. Corpus of Persian language, compiled byShlomoArgamon, a research group, at Illinois Institute of Technology,has been used. To process the corpus data, Sketch Enginesoftware has been used. The list of core Persian vocabulary is supposed to be useful for Persian language teachers, learners,lexicologists, lexicographers and Persian grammarians.

Keywords: *Persian, language learning, vocabulary building, corpus, word list.*

Introduction:

The role of vocabulary in learning a foreign language is an established phenomenon. According to Mukoroli(2011) vocabulary constitutes the integral procedure to learn a foreign language. Vocabulary of any language has been taught traditionally with various techniques and strategies. All these vocabulary building techniques have been originally teacher centered and the learners have been at the lowest paradigm of autonomy. The results of this orthodox approach to foreign language learning were not motivating. Moreover, with the emergence of new approach of lexical item, the range of vocabulary of a language has also increased many times. The lexicon of a language ranges from single words to compounds, idioms, collocations, etc. for example, **چرخ** (single word), طبقه بندی(compound), لجشگرفته (idiom), مشروبقوی (collocation), etc.

There has been felt a dire need to improve the situation. With the dawn of 21st century and the discovery of computers and language corpora, the situation has started improving to meet the vocabulary needs of the foreign language learners. Schimitt (1997) describes five types of vocabulary learning strategies: Social strategy, Metacognitive strategy, Memory strategy, Determination strategy and Cognitive strategy. Tradition ally, learners have been using Determination and Cognitive strategies that are comparatively complex and have not proved effective. In comparison with Schimitt's (1997) strategies of vocabulary learning, Qian (2002) has extended the vocabulary issue and re-divided the foreign vocabulary in terms of size, depth, lexical organization, and autonomy of the learners.

*Department of English University of Sahiwal Pakistan

**SLCP, Universiti Utara Malaysia

***Department of English University of Sargodha, Pakistan

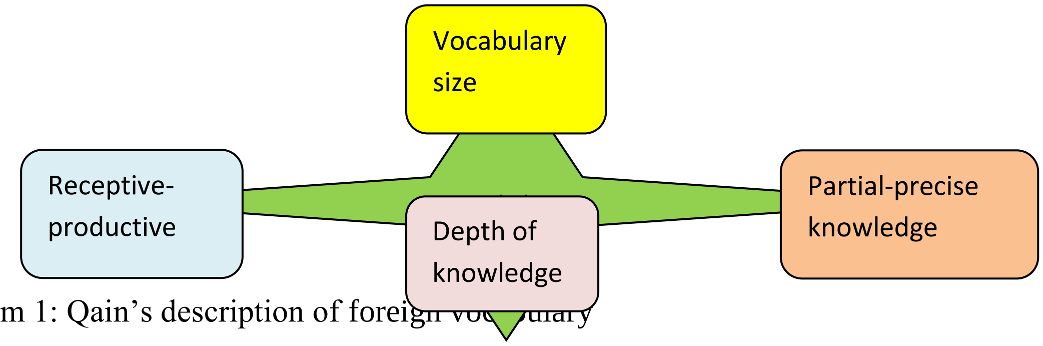

Diagramm 1: Qain's description of foreign vocabulary

According to Adger (2002) vocabulary items cannot be analyzed only from semantic perspective.They cover the structural aspect of language as well. The vocabulary items motivate the learners to build the possible combinations as well. He considers vocabulary building a continuous process with significant comprehension of the context as well.

The use of language corpora has further improved the situation. It has altogether changed the trends in foreign language teaching. In this regard, frequency word list with the help of corpus have proved very useful to enhance the overall proficiency of the learners.

Statement of the Problem

Persian is one of the oldest languages of the world with more than 115 million speakers currently. Apart from newly acquired geo-political and geo-economical position, Persian has traditionally been a source of rich culture and civilization. Moreover, Persian has strong roots with Islamic history and has been the dominating political language of the subcontinent. "Language is culture" is an established phenomenon. Similarly, Persian language reflects cultural roots, ideologies, values and the sentiments of the speakers.It is taught as a foreign language in Pakistan at various levels. The undertaken research basically highlights core Persian vocabulary for the Pakistani learners of the Persian language. The research will help the Pakistani Persian learners to handle language resources successfully and effectively.

Research Methodology

Corpus has revolutionized every aspect of foreign language learning. Corpus is a large collection of electronically collected text. The text is used for various types of analyses and inferences. A good corpus contains text from all spheres of life. The Talkbank Persian corpus is originally a corpus of Persian language that has been compiled of blog posts from different Persian blog sites. ShlomoArgamon, a research group, has collected the text to compile the Persian corpus at Illinois Institute of Technology. The Corpus is constantly being upgraded and more and more files are being added due to the fact that that the larger corpus, the better the analysis is. In order to process corpus data, Sketch Engine has been used. Sketch Engine is remarkable software for statistical analysis. It performs various functions as word sketch, word lists, n-grams, concordance, and thesaurus. For the undertaken research, sketch engine has been used to retrieve word list from the Persian corpus.

Persian Language and the Technology

There have been incessant innovations in foreign language teaching across the globe. Persian language is not an exception to this change. Institutions of Persian language teaching and learning are adopting themselves with the advent of new technological techniques to make the

Persian language process easy for the learners. Focus has been on the development of curriculum, instruction and assessment of Persian language according to the needs of the hour. It has been realized that computer based tools have played a significant role in handling Persian language learning for non-native speakers.Rather, those tools are more motivating than the traditional methods of learning Persian language.With the advent of computers, it has also placed the Persian language learners in a new role. Similarly, the role of the teachers of the Persian language has also been revolutionized. Teacher’s role is more of a facilitator rather than a police man of the classroom. Learning of the language occurs in a co-operative manner, as it is termed as “co-operative learning” in foreign language learning.

The techniques and strategies to teach Persian language to non-native speakers of Persian language need to be modernized. The retrieved word list can be considered to execute the prospective change in teaching and learning of Persian as a foreign language. This is a new scenario at least, for the Pakistani learners of Persian language.It will provide strong foundations to learn the core Persian vocabulary.The list may be further divided according to the level of learners. Non-native learners of Persian language will have to learn only those Persian words that are in use by the native speakers, thus, preparing the learners to perform in the real Persian language context, the language that that is used by Persian native speakers in their daily life. The wide range of the corpus files means that the range of the Persian vocabulary is also very extended e.g. health, exercise, sports, politics, social issues, arts, media, etc. For example, the corpus has determined the frequencyof the Persian parts of speechthat can help learners and teachers of Persian language to determine the focus of their focus of linguistic activity.

Tag	**Probability**
COMmon Noun	40%
PRoper Noun	17%
SIMpleADJective	26%
Verb	2%
RESidual	10%
Others	5%

word(512,643 items | 471,371,941 total frequency)

Word ⍰Frequency Per Million

1 و	20,366,133	42,897
2 در	14,292,826	30,105
3 بهه	13,337,645	28,093
4 از	11,097,460	23,374
5 کهه	9,190,561	19,358
6 را	7,397,265	15,581
7 این	6,712,046	14,137
8 با	5,934,386	12,499
9 است	5,849,732	12,321

Word ⍰Frequency Per Million

51 پس	660,356	1,391
52 دست	642,200	1,353
53 نیست	640,116	1,348
54 بازدید	634,623	1,337
55 امریکا	624,850	1,316
56 دولت	606,883	1,278
57 کند	597,041	1,258
58 مورد	593,426	1,250
59 شود	591,735	1,246

Word ⍰Frequency Per Million

101 حضور	396,685	836
102 زندگی	392,565	827
103 امام	389,581	821
104 میکنند	385,858	813
105 قبل	382,033	805
106 اعلام	381,796	804
107 شدن	381,664	804
108 زمان	381,467	803
109 کردن	381,320	803

Word ⍰Frequency Per Million

151 شرکت	302,479	637
152 ظریف	302,198	637
153 تولید	301,749	636
154 بخش	299,455	631
155 زیر	299,221	630
156 اب	295,113	622
157 بودند	294,535	620
158 بیش	290,051	611
159 چون	289,977	611

Vol.4 No.3 2021

ISSN Online: 2709-7625

ISSN Print: 2709-7617

برای10	2,881,488	6,069
یک11	2,153,287	4,535
ان12	2,055,236	4,329
ایران13	1,901,952	4,006
خود14	1,803,021	3,798
هم15	1,783,698	3,757
بود16	1,733,332	3,651
شده17	1,650,113	3,476
تا18	1,646,097	3,467
شد19	1,559,922	3,286
بر20	1,557,599	3,281
کرد21	1,543,963	3,252
من22	1,393,958	2,936
نظر23	1,364,880	2,875
ما24	1,139,929	2,401
سال25	1,079,621	2,274
یا26	1,045,206	2,201
گفت27	1,035,052	2,180
اما28	1,008,835	2,125
او29	978,677	2,061
میشود30	946,418	1,993
باید31	943,770	1,988
هر32	917,714	1,933
نیز33	887,099	1,868
دارد34	863,913	1,820
کشور35	848,220	1,787
روز36	833,101	1,755
دیگر37	798,801	1,682

بعد60	585,973	1,234
میکند61	577,975	1,217
رو62	563,415	1,187
یکی63	562,178	1,184
اسلامی64	556,726	1,173
اینکه65	537,948	1,133
روی66	532,640	1,122
استفاده67	532,507	1,122
داد68	515,876	1,087
تهران69	506,996	1,068
وجود70	502,138	1,058
خواهد71	495,307	1,043
رییس72	493,320	1,039
حال73	489,414	1,031
ادامه74	487,880	1,028
وزیر75	480,953	1,013
هستند76	470,499	991
پیش77	467,578	985
نه78	467,015	984
داشت79	462,036	973
مجلس80	461,880	973
عنوان81	459,488	968
خبر82	458,027	965
بین83	454,860	958
چند84	451,036	950
بیشتر85	445,272	938
صورت86	442,753	933
انجام87	442,131	931

بوده110	380,493	801
توسط111	378,703	798
سر112	377,181	794
سهم113	376,379	793
جدید114	375,401	791
پاسخ115	373,624	787
تیم116	371,571	783
تصاویر117	363,185	765
توجه118	361,483	761
مذاکرات119	361,247	761
امروز120	360,942	760
راه121	352,002	741
کردند122	349,696	737
نام123	347,714	732
تنها124	344,601	726
انقلاب125	344,001	725
وقتی126	343,611	724
ملی127	342,870	722
اخرین128	342,849	722
اول129	341,572	719
حقوق130	341,001	718
دارند131	340,160	716
بسیار132	335,576	707
اقای133	333,964	703
بازی134	330,346	696
گذشته135	327,298	689
افزایش136	326,333	687
برخی137	325,861	686

الله160	288,466	608
میان161	288,340	607
طرح162	288,273	607
گرفته163	288,168	607
جمهوری 164	287,024	605
ایا165	286,350	603
وزارت166	285,736	602
بار167	285,348	601
نفر168	285,323	601
اخبار169	285,275	601
انتشار170	283,007	596
دلیل171	281,541	593
بدون172	278,982	588
سایت173	277,924	585
یه174	277,023	583
باز175	276,670	583
اشاره176	276,630	583
قیمت177	274,601	578
انتخابات178	274,506	578
افراد179	274,486	578
مطالب180	274,262	578
پایان181	272,738	574
نشان182	271,743	572
فقط183	269,304	567
شورای184	265,392	559
بزرگ185	264,974	558
روحانی186	264,378	557
برنامه187	263,742	556

38چه	785,991	1,656
39همه	785,037	1,653
40مردم	781,802	1,647

88کنید	440,890	929
89ولی	427,436	900
90تمام	426,078	897

138ماه	325,795	686
139علی	321,036	676
140سازمان	316,070	666

188دانشگاه	261,818	551
189همچنین	259,377	546
190ابان	259,067	546

41وی	749,229	1,578
42کار	744,909	1,569
43اگر	736,969	1,552
44کرده	725,346	1,528
45شما	717,357	1,511
46دو	711,052	1,498
47تو	701,165	1,477
48باشد	683,907	1,440
49انها	683,438	1,440
50قرار	668,673	1,408

91هیچ	425,484	896
92سیاسی	418,921	882
93کنند	417,044	878
94عکس	412,993	870
95داشته	412,648	869
96گزارش	408,364	860
97همین	405,010	853
98تاریخ	399,207	841
99چرا	398,765	840
100درباره	397,715	838

141داده	315,730	665
142جهان	315,484	664
143کسی	315,379	664
144خیلی	313,457	660
145هزار	311,329	656
146حتی	310,434	654
147شهر	307,631	648
148ساعت	307,043	647
149فیلم	303,822	640
150جامعه	303,038	638

191سوی	258,5
192ایجاد	256,2
193خدا	254,8
194ندارد	254,3
195کاهش	253,9
196قانون	253,6
197علیه	252,9
198اجتماعی	250,3
199معاون	249,9
200ایرانی	249,7

Conclusion

Persian vocabulary has been highlighted as the most critical indicator in learning Persian as a foreign language; however, at the same time, it has also been recognized as one of the challenging areas of language learning. Luckily, the computer as language tool has decreased the pressure from the learners and the teachers of the Persian language. The retrieved frequency list of Persian words is supposed to be very useful for the non-native learners of the Persian language.Rather, the learners of Persian language may be motivated to use the Persian corpus in a more productive manner. The undertaken research is an innovation of its sort and fills a critical crevice in research on Persian vocabulary.

Recommendations:

The undertaken research, by no means, is complete in its entirety but it can be used as appoint of departure. The retrieved Persian word list may be made more focused and precise by extending the base volume of the Persian corpus. However, the research can motivate the Persian lexicologists to further investigate the various dimensions of Persian vocabulary.

References

Akhter, S. (2020). Towards Cultural Clash and Hybridity, An Analysis of Bapsi Sidhwa's An American Brat. *sjesr*, *3*(3), 22-34.

Akhter, S., Ajmal, M., & Keezhatta, M. S. (2020). A case study on the effectiveness of learner autonomy in English literature classroom. *PalArch's Journal of Archaeology of Egypt/Egyptology*, *17*(6), 3063-3076.

Akhter, S., Haidov, R., Rana, A. M., & Qureshi, A. H. (2020). Exploring the significance of speaking skill for EFL learners. *PalArch's Journal of Archaeology of Egypt/Egyptology, 17*(9), 6019-6030.

Akhter, S., Kausar, R., & Faisal, M. (2020). Towards the description of newspapers in learning English. *International Journal of Management (IJM), 11*(9).

Coniam, D. (1997) 'A preliminary inquiry into using corpus word frequency data in the automatic generation of English language cloze tests'. CALICO Journal 16/2-4, 15-33.

Granger, S. (1998) The computer learner corpus: a versatile new source of data for SLA research. 3-18. In S. Granger (ed.) Learner English on Computer. London: Longman.

Henriksen, B. (1999). Three dimensions of vocabulary development. Studies in SecondLanguage Acquisition, 21(2), 303–317. http://dx.doi.org/ 10.1017/S0272263199002089

Hoey, M. (2000) 'A world beyond collocation: new perspectives on vocabulary teaching' in M. Lewis (ed.) Teaching Collocations, pp. 224-245.

Liu, L., Akhter, S., & Qureshi, A. H. (2020). Towards the Description of Techniques in Teaching L2 Vocabulary. *Revista Argentina de Clínica Psicológica, 29*(3), 268.

Mukoroli, J. (2011).Effective Vocabulary Teaching Strategies for The English For Academic Purposes EslClassroom.MATESOLCollection.Paper 501.

Moore, M., & Calvert, S. (2000). Brief report: Vocabulary acquisition for children with autism: Teacher or computer instruction. Journal of Autism and Developmental Disorders,30(4).

Neuman, S. B., & Dwyer, J. (2009). Missing in action: Vocabulary instruction in pre-K. The Reading Teacher, 62(5), 384–392.

Qian, D.D. (2002). Investigating the Relationship between Vocabulary Knowledge and Academic Reading Performance: an assessment perspective. *Language Learning,* 52(3), 513-536

Schmitt, N. (1997). Vocabulary Learning Stratigies.Vocabulary, Description, Acquisition and Pedagogy.Pp.199-228. Cambridge University press