STANISLAV MARTYNYUK

Statistical Approach to the Debate on Urdu and Hindi

Recent debate on the relationship between Urdu and Hindi has mostly addressed cultural and cognitive differences. These differences have been observed on different cognitive verbal and non-verbal tasks and conclusions on the variability in language use and comprehension have been drawn from their results. In this paper, a different approach is presented. The pilot study which is described below examines statistical lexical distribution in electronic media speech. Tools for comparison of lexical data were developed to compile frequency lists for Urdu and Hindi. The lists were analyzed using a cross validation approach based on word frequencies. On the basis of these measurements, statistical analysis is able to establish the percentage difference between frequent words used in the two languages. The results are discussed in detail including correlation and principal components analysis. The paper concludes that there are interesting differences across the two languages on the measures studied and further research in this area is warranted.

A number of papers published in previous issues of the *AUS* addressed the emergence of Urdu and Hindi. In his article entitled "Some Notes on Hindi and Urdu" (No. 11), Ralph Russell agrees that the *one language, two scripts* approach to the issue is far from being true and natural. He analyzes several publications on Urdu and Hindi and the vocabulary presented in those publications. Russell concludes that Urdu and Hindi are two separate languages and he suggests that they be treated as such "despite their almost completely common structure and less completely common stock of everyday words" (p. 204). In his paper entitled "Urdu in India" (No. 17), David Matthews says that the relationship between Urdu and Hindi is "extremely complex, and in arguments dealing with the sensitive issue of language the case is usually grossly oversimplified.” He adds:

> At the most basic level Hindi and Urdu, leaving aside their scripts, are virtually identical languages and serve admirably, as they always have done, as a valuable link between all South Asian communities wherever they may reside. At certain levels they are very different from each other and deserve separate treatment and study. (p. 157)

Matthews concludes that, considering the directions in which the two languages are going in India and Pakistan, with the course of time the two languages “will inevitably drift apart and there will be even less common ground between them than there is at present” (*ibid.)*. Further Matthews draws a parallel with the linguistic situation in Ukraine, a former Soviet republic. However he is mistaken in saying that Ukrainian is “very closely related” to Russian. Recent research indicates that the linguistic difference between the vocabularies of Ukrainian and Russian is about 38 percent.[1] Notable is the fact that the difference between Russian and, for instance, Serbo-Croatian is 36 percent[2] yet no one has ever suggested that these two languages are closely related. Also, the lexical distance between, for instance, Italian and French is 30 percent, and between Spanish and Romanian is 57 percent (*ibid.*). According to Tyschenko, the difference between English and German is 49 percent, and the difference between Ukrainian and Polish is 30 percent.

The above-mentioned qualitative taxonomy data for European languages was obtained as a result of comparative analyses of the most frequent words used in those languages. The researchers employed comparative-historical methods that are based on universal laws. Modern science recognizes dynamic and statistical laws. A dynamic law reflects the dependence between separate states of an object, each of which predetermines the following one. A statistical law reveals the objective dependence between the batteries of similar and relatively independent things.[3] Linguistic theory has had the knowledge of a number of objective statistical

[1] Kostiantyn Tyschenko, *Metateoriya Movoznavstva* (Kyiv: Osnovy, 2000), p. 265.

[2] A. Shaikevich, *Gipotezy o Estestvennykh Klassakh i Vozmozhnost Kolichstvennoi Taksonimii v Lingvistike // Gipoteza v Sovremennoi Lingvistike* (Moscow: Nauka, 1980), p. 331.

[3] *Filosofskiy Slovar* (Kyiv: Naukova Dumka, 1973), p. 497.

laws and regularities which act in speech without reference to the will and awareness of speakers. These laws were discovered by stenographer Jean-Baptiste Estoup and philologist George K. Zipf, whose works were later deepened and expanded by linguists and mathematicians such as J. C. Willis, G. U. Yule, Benoit B. Maldelbrot, Gustav Herdan, S. C. Bradford, M. V. Arapov, M. M. Kherts and others. Observations of frequency dictionaries conducted by Zipf as early as the 1930s demonstrated that the use of words by people in speech is governed by a drive for optimizing the relationship between the requirement of diversity and the speaker's tendency to exert the least effort.[4]

Before I move on to the description of the research I did which employed a very simple but rather old and effective approach to comparing languages, I would like to step briefly into my personal history in Urdu studies. I began to learn Urdu in 1997, having studied Hindi at the Kyiv National Taras Shevchenko University in Ukraine for two years. At first, really embarrassing for me was the close relationship between the two languages in terms of grammar and pronunciation on the one hand, and the "strange" need to select proper words for Urdu speech which would not sound Hindi, as was demanded by our Urdu lecturer who was a Pakistani, on the other. Further on, my interest in finding the borderline which would divide the vocabularies of the two languages grew to the extent that I decided to undertake research into the lexical differentiation between Urdu and Hindi.

To compare Hindi and Urdu I decided to put aside the visual and audible differences between the two languages and employ a comparative method to solve the problem. I have already mentioned above that statistics can help establish the lexical difference between languages. Through my personal observation I have come to the conclusion that in bilingual societies, like the one in Ukraine, speakers do not recognize the real difference between the two languages. In order to persuade speakers in Ukraine that Ukrainian and Russian are not as closely related as they are believed to be, I had to cite the statistics. In Ukraine, almost all the people have no difficulty understanding Russian, whether in everyday situations or in television and radio broadcasts. In other words, if a person speaks Russian, he/she will have no difficulty in communicating. However, the situation will be quite different for Russian speakers coming

[4] George K. Zipf, *Human Behavior and the Principle of Least Effort: An Introduction to Human Ecology* (New York: Hafner, 1965).

to Ukraine from abroad. If they have never lived in Ukraine for an extended period or studied Ukrainian, they will have great difficulty understanding Ukrainian, even in ordinary speech, not to mention the language used in the media.

I think that the linguistic situation in Hindi and Urdu speaking areas is similar to the one described above to some extent. Speakers in bilingual or monolingual societies do not recognize the real differences in the use of words in speech. It is a common assumption of non-linguists in India and Pakistan that the difference between Hindi and Urdu mostly lies in the use of different alphabets and the selection of proper words, i.e., Sanskrit-derived words for Hindi and Arabic and Persian words for Urdu.

The approach which was employed in my research is based on the statistical finding that different speech samples, including electronic media speech, follow a set of statistical laws, e.g., Zipf's law (1932), which makes them comparable on the basis of word frequencies. Zipf's law, named after the Harvard linguistics professor George Kingsley Zipf (1902-1950), is the observation that the frequency of occurrence of some event (*P*), as a function of the rank (*i*) when the rank is determined by the above frequency of occurrence, is a power-law function $P_i \sim 1/i^a$ with the exponent *a* close to unity.[5] The most famous example of Zipf's law is the frequency of occurrence of words in a language. The formula for Zipf's law here is $rf = C$ where r is the rank of a word, f is the frequency of occurrence of the word, and C is a constant that depends on the text being analyzed.

Any two languages can be compared by contrasting the words that constitute their vocabularies. The issue crucial at this point is to select the right words so that the general picture will not be distorted. The most appropriate idea in this context is to compare the most frequent words used in both languages.

The universal statistic law, which was discovered by Zipf in 1932, is famous for its practical applications. For instance, according to Lidia Zasorina, by learning the 230 most frequent words of a language a student will be able to "recognize" one half of the words in any text; the 1000 most frequent words, two thirds of a text; and the 5000 most frequent words, 90 percent of a text.[6] According to the Zipf-Guiraud law (1959), to

[5] Wentian Li, *Zipf's Law* <http://linkage.rockefeller.edu/wli/zipf/>.

[6] L. Zasorina, ed., *Chastotny Slovar Russkogo Yazyka* (Moscow: Russkiy Yazyk, 1977), pp. 895-915.

understand one half of any text a person needs to know the 1000 most frequent words, and for understanding two thirds of any text about 4000 of the most frequent words.[7]

The most recent frequency vocabulary I am aware of is the *Frequency Dictionary for Russian* that was compiled by Serge Sharoff in 2002.[8] According to Sharoff, the 1000 most frequent lemmas cover 64.1 percent of word forms in texts, the 2000 most frequent lemmas cover 72 percent of word forms in texts, the 3000 most frequent cover 76.7 percent, and the 5000 most frequent cover 82.1 percent. While Zasorina's dictionary is relatively small by modern standards, about 1 million words, the dictionary compiled by Sharoff is based on a 40 million word corpus. Statistical regularities are the basis of the structure of the vocabulary of any language or text. Zipf's law is a reflection of a specific property of the organization of human memory, which usually operates with more frequent language units in all cases of the spontaneous use of speech.

But let us now turn to Urdu and Hindi. It is obvious that the above mentioned laws can and should be used in the comparative analysis of the vocabularies of these two languages. The only thing we need is to have an instrument to work with. It is remarkable how little serious attention seems to have been paid to this task. As far as I know, there has been no comparison done between the two languages which was based on frequency lists. Since I was unable to find frequency dictionaries for Hindi and Urdu, though they may possibly exist, I decided to compile such lists on my own. At first, I wanted to compile frequency dictionaries based on modern Hindi and Urdu. I had hoped to take samples for the compilation of corpora from the Internet. However, I very soon discovered that Hindi and, even more so, Urdu resources on the Internet are limited. This led me to the conclusion that my research should be narrowed.

It's obvious that to count words in a text one needs to have appropriate tools. The main tool for my research was a computer application specifically designed to count the number of occurrences of every word form in a text. By processing the text through this application a list of all word forms found in the text is produced, arranged in descending order of the frequency of the occurrences of each form in the text. In the case of

[7] M. Lehner, *Der Englische Grundwortschatz* (Leipzig: Veb Enzyklopaedie, 1975), S.6.

[8] Serge Sharoff, *Frequency Dictionary for Russian*, 2002 <http://www.artint.ru/projects/frqlist/frqlist-en.asp>.

Urdu and Hindi, the application only allowed for the recognition of word forms which I later had to manually bring together to lemmas (word forms used in dictionaries) using the MS Excel application. Since words appear in different forms in speech, I had to prevent them from being counted as separate words by means of *lemmatization*:

1. Verb forms were reduced to the infinitive.
2. Inflected forms of nouns were reduced to the nominative singular.
3. Inflected forms of adjectives were reduced to the nominative masculine singular.
4. Comparatives and superlatives of gradable adjectives were reduced to the absolute form.

All this was done manually with the help of the MS Excel application. For instance:

		freq		freq		freq		freq
Lemma:	aana		naya		rukn		party	
	aa	51	nae	195	rukn	77	parties	27
	aae	124	naee	174	arkaan	261	partiyoN	44
	aaee	46	naya	71	Total	**338**	party	548
	aaee	40	Total	**440**			Total	**619**
	aaeN	26						
	aakar	31						
	aane	173						
	aaya	78						
	Total	**569**						

Only lemmas were included in the frequency lists, accompanied by the total frequency of the occurrence of a word.

The frequency lists for both Urdu and Hindi are based on corpora of about 440,000 words each. Since the application was only capable of counting words in one text, I had to collect a great number of small pieces into a corpus. Another technical issue at that stage was that all the pieces of the text in each corpus had to be encoded in the same format. Standards for encoding large lexical resources are under active development now, not only for Urdu and Hindi but also for other languages. However, since the original aim of my research required the availability of a large corpus for each of the two languages, this probably was the main reason I had to narrow my research. Since texts had to be encoded in one font, I needed to take the pieces from sites that used that font.

For Hindi, I decided to copy texts from *webdunia.com*. This site uses the namesake Hindi font which can be freely downloaded. It has an existing news service which is divided into three major categories: international, national, and regional. The corpus consists of Hindi news feeds taken from the three categories between July and September 2002. With Urdu, however, the issue was much more complicated. As is known, Urdu readers in Pakistan prefer to read newspapers published in a Nastaliq font. Many Internet-based Pakistani newspapers use, for instance, the InPage Noori Nastaliq system of Urdu calligraphy. However, the biggest problem here is that almost all the news feeds on Pakistani websites are published in a scanned version, which made them impossible to use in computer processing. I failed to find a Pakistani site which used encoded text. Finally, I decided the best solution was to use romanized Urdu news feeds. News in this format is published by one of Pakistan's largest newspapers, the daily *Jang* (*jang.net*). Its roman-letter Urdu news service is also divided into three categories: international, major (national), and regional. To obtain the approximately 440,000 words needed, I had to take all of the romanized Urdu news feeds published between May and November 2002.

By processing each corpus with the word counting computer application I was able to produce unlemmatized frequency lists. After the manual lemmatizing was completed it became possible to cross-check the words in the two lists. The 100 most frequent words for Hindi and Urdu used in the electronic media are as follows:

Table 1[9]

	HINDI		URDU	
word rank	frequency	word	frequency	word
1	43098	ka	38141	ka
2	24149	hona	21514	hona
3	15141	meN	13718	meN
4	10553	ne	12433	karna
5	9346	karna	9705	ne
6	9220	ko	9157	aur

[9] Urdu and Hindi words that appear in this and the next table have been left untransliterated. They are spelled well enough to be easily recognized. —*Editor*

7	8892	se	8586	se
8	8134	jana	7745	ko
9	8029	ki	7681	jana
10	7329	yah	6289	par
11	6614	aur	6118	keh
12	5168	ve	4680	dena
13	4896	par	4611	yah
14	4545	kahna	3841	kahna
15	4019	dena	3294	voh
16	3178	bhee	2908	(ke)liyay
17	3106	rahna	2479	naheeN
18	3003	nahiN	2421	ek
19	2862	ek	2018	rahna
20	2661	(ke) lie	1883	jo
21	2135	vah	1748	sadar
22	1825	lena	1729	bhee
23	1630	hee	1673	lena
24	1586	apna	1390	Amreekee
25	1457	chunaav	1315	hukoomat
26	1416	sarkaar	1266	Bhaaratee
27	1374	batana	1231	koee
28	1368	koee	1184	mulk
29	1324	baad	1183	apna
30	1285	jo	1168	mutaabiq
31	1181	parti	1168	afraad
32	1149	sakna	1149	halaak
33	1100	rajya	1111	fauj
34	1079	ye	1017	hamlah
35	1072	desh	989	sakna
36	1031	sath	981	ba‘d
37	1016	log	966	khilaaf
38	951	lekin	890	2

39	933	tak	774	general
40	927	pahla	745	police
41	912	rashtrpati	736	tak
42	902	atankvadee	734	jaaree
43	888	maamla	698	ijlaas
44	863	tatha	674	faislah
45	862	neta	672	intekhaabaat
46	856	ana	652	‘ilaaqah
47	815	to	643	jabkeh
48	801	chahna	627	3
49	775	pulis	624	election
50	747	do	619	party
51	710	baat	611	faujee
52	707	banana	610	qaumee
53	695	beech	608	zakhmee
54	686	adhikari	599	rakhna
55	631	yahaN	569	aana
56	615	dvara	550	arab
57	615	kaNgres	549	to
58	603	karana	542	ham
59	601	pradhanmaNtri	538	banaana
60	583	dal	534	bataana
61	571	milna	531	assembly
62	560	varsh	528	ghair
63	550	adhyaksha	527	saath
64	548	baare	526	shuroo‘
65	546	maNtri	487	dauraan
66	544	hamla	481	dehshat gardee
67	543	aaj	476	saal
68	540	kaaran	476	muzaakraat
69	536	baithak	470	siyaasee
70	526	anya	468	dhamaakah

71	523	aisa	466	tamaam
72	521	amerikee	463	hee
73	520	teen	450	darmiyaan
74	518	purv	450	muslim
75	517	mukhyamaNtri	448	baat
76	514	dauran	447	4
77	510	anusaar	446	rupe
78	510	kuchh	446	5
79	506	tarah	442	donoN
80	496	sabhee	441	samet
81	495	suraksha	440	naya
82	494	poora	431	haal
83	493	maarna	426	taur
84	492	samay	424	Pakistani
85	490	bharatiy	421	jang
86	484	sootra	417	gariftaar
87	483	ab	414	sarhad
88	480	khilaaf	392	kasheedgee
89	478	baatcheet	389	Israeli
90	477	naya	382	court
91	458	sadasya	376	mazeed
92	457	jaaree	372	khitaab
93	451	sthiti	372	league
94	442	din	371	jamaa'at
95	442	roop	368	chief
96	439	donoN	366	kaam
97	434	shaamil	366	chaahna
98	428	aarop	362	taaham
99	425	kam	362	firing
100	420	shuroo	357	kam
101	418	kshetr	355	tarjumaan
102	414	saNgathan	352	commission

Total	245342		223272	

Note: 1) words marked in green matched fully; 2) words marked in blue did not match; 3) words marked in yellow mean names of institutions and were not compared; 4) words marked in light blue partially matched but were not included in the results; 5) words not marked were not compared. Names of people and geographical places were not included in the lists.

The first 100 words in the Hindi list constitute 55.61 percent of the Hindi corpus which was made up of 441,153 words. The first 100 words in the Urdu list account for 50.64 percent of the Urdu corpus which included 440,929 words.

Since the frequencies of words which mean the same things are different in Hindi and Urdu, I chose to make the comparison on the basis of the Hindi frequency list. Words were taken from the Hindi list one by one from the very top to the bottom and equivalents for them were found in the Urdu list. It may be seen from Table 1 that the ranks of the first three words in Hindi and Urdu are the same. Further, words in the Hindi list matched fully with their equivalents in Urdu until the word ranked twenty-fifth. It can also be seen from Table 1 that some words of Persian and Arabic origin are used both in Hindi and Urdu, such as *maamla*, *hamla*, *khilaaf*, etc.

The comparison results obtained for words which mean the same thing are arranged in the following table:

Table 2

HINDI			URDU		
word rank	frequency	word	word rank	frequency	word
25	1457	chunav	45	672	intekhaabaat
			49	624	election
26	1416	sarkar	25	1315	hukoomat
33	1100	rajya	52	610	qaumee
			138	287	qaum
			170	239	soobah
			349	127	riyaasat
35	1072	desh	28	1184	mulk
37	1016	log	31	1168	afraad
41	912	rashtrpati	21	1812	sadar
42	902	atankvadee	182	226	dehshat gard
45	862	neta	106	344	rahnuma

			293	145	leader
53	695	beech	73	450	darmiyaan
54	686	adhikaree	153	262	sarkaaree
			1010	50	‘ohdedaar
56	615	dvara	730	66	zarie‘
59	601	pradhan-mantri	137	292	vazeer-e-aa‘zam
60	583	dal	94	371	jamaa‘at
			226	178	group
62	560	varsh	67	476	saal
63	550	adhyakshya	192	211	chairman
65	546	mantri	640	75	vazeer
68	540	karan	225	178	vajah
69	536	baithak	43	698	ijlaas
70	526	anya	62	528	ghair
74	518	purv	187	214	saabiq
			911	54	mashriqee
75	517	mukhya-mantri	1854	24	vazeer-e-a‘ala
77	510	anusar	30	1168	mutaabiq
80	496	sabhi	71	466	tamaam
81	495	suraksha	186	219	security
84	492	samay	171	238	waqt
86	484	sutr	131	301	zaraae‘
91	458	sadasya	111	338	rukn
93	451	sthiti	82	431	haal
94	442	din	130	302	roz
95	442	rup	581	83	surat
98	428	arop	126	310	ilzaam
101	418	kshetr	46	652	‘ilaaqah
102	414	sangathan	207	190	tanzeem

Note: Some of the words in the Hindi list have several meanings for which Urdu equivalents were found. Such Urdu words are framed.

It can be seen from Table 2 that a total of 34 words in the list of the 100 (102) most frequent Hindi words did not match with Urdu words of the same meaning and had to be translated.

Conclusion

The research reported in this paper led to several conclusions. First, we may conclude that one third of the 100 most frequent words used in Hindi and Urdu are different. We must remember, however, that the corpora were compiled on the basis of news feeds that appeared in the electronic media, therefore this conclusion cannot be taken as evidence that the lexical distance between the modern Hindi and Urdu languages is about 30 percent. Still, it does suggest that further research may be warranted to establish the level of differentiation between Hindi and Urdu. Second, the approach used in this pilot study is useful since the two languages can be easily compared on the basis of word frequencies.

Third, the frequency lists that were obtained can be used for the compilation of bilingual dictionaries. Also, the lists can and should be used in teaching Hindi and Urdu, especially in the development of student books. As has been mentioned above, by learning the most frequent words students can greatly reduce the time needed for achieving proficiency in either language. Another important contribution of this research is that corpora for Hindi and Urdu have now been compiled in an electronic format.

In this paper I have tried to briefly present the details of one method for the comparison of the vocabularies of Hindi and Urdu as used in the electronic media. Although I had very limited funds and had to do everything completely on my own, I hope that my findings will help arouse interest among linguists for conducting further research in this area. ❒