10

1965 International Conference on
Computational Linguistics

MEASUREMENT OF SIMILARITY BETWEEN NOUNS

Kenneth E. Harper

The RAND Corporation
1700 Main Street
Santa Monica, California 90406

ABSTRACT

A study was made of the degree of similarity between pairs of Russian nouns, as expressed by their tendency to occur in sentences with identical words in identical syntactic relationships. A similarity matrix was prepared for forty nouns; for each pair of nouns the number of shared (i) adjective dependents, (ii) noun dependents, and (iii) noun governors was automatically retrieved from machine-processed text. The similarity coefficient for each pair was determined as the ratio of the total of such shared words to the product of the frequencies of the two nouns in the text. The 780 pairs were ranked according to this coefficient. The text comprised 120,000 running words of physics text processed at The RAND Corporation; the frequencies of occurrence of the forty nouns in this text ranged from 42 to 328.

The results suggest that the sample of text is of sufficient size to be useful for the intended purpose. Many noun pairs with similar properties (synonymy, antonymy, derivation from distributionally similar verbs, etc.) are characterized by high similarity coefficients; the converse is not observed. The relevance of various syntactic relationships as criteria for measurement is discussed.

MEASUREMENT OF SIMILARITY BETWEEN NOUNS

1. INTRODUCTION

One of the goals of studies in Distributional Semantics is the establishment of word classes on the basis of the observed behavior of words in written texts. A convenient and significant way of discussing "behavior" of words is in terms of syntactic relationship. At the outset, in fact, it is necessary that we treat a word in terms of its Syntactically Related Words (SRW). In a given text, each word bears a given syntactic relationship to a finite number of other words; e.g., a finite number of words (nouns and pronouns) appear as "subject" for each active verb; another group of nouns and pronouns are used as "direct object" of each transitive verb; other words of the class, "adverb," appear as modifiers of a given verb. In each instance we may speak of the related words as SRW of a given verb, so that in our example three different types of SRW emerge; a given SRW is then defined in terms both of word class and specific relationship to the verb. (A given noun may of course belong to two different types of SRW, e.g., as both subject and object of the same verb.)

Distributionally, we may compare two verbs in terms of their SRW. The objective of the present study is to test the premise that "similar" words tend to have the same SRW. This premise is tested, not with verbs, as in the

above example, but with nouns. Our procedure is (1) to find in a given text three types of SRW for a small group of nouns, (2) to find the number of SRW shared by each pair of nouns formed from the group, and (3) to express the "similarity" between individual nouns, and groups of nouns, as a function of their shared SRW. Another example: it might turn out that in a given text the nouns "a" and "b" ("avocado" and "cherry") share such adjective modifiers as "ripe," whereas nouns "c" and "d" ("chair" and "furniture") have in common the adjective modifier "modern." These facts would lead us to conclude that "a" and "b" are similar, that "c" and "d" are similar, that "a" and "c" are less similar, etc.

A number of questions arise: What is "similarity" anyway? Do words that are similar in meaning really share a significant number of SRW in a given text? What is "a significant number"? Do not dissimilar words also have many common SRW? How much text is necessary in order to establish patterns of word behavior? What is the effect of multiple-meaning in words, and of using texts from different subject areas? The present investigation should be regarded as an experiment designed to throw some light on these questions; no validity is claimed for the "results" obtained. Our audacity in attempting the experiment at all is based on three factors: the possession of a text in a limited field (physics), the foreknowledge that the multiple-

meaning problem is minimal, and the capability for automatic processing of text. (The latter is clearly a necessity, in view of the size and complexity of the problem.) The reader may well conclude that the experiment proves nothing. We would hope, however, that such an opinion would not preclude a critical judgment of the procedures employed, or the suspension of disbelief if the results do not correspond with his expectations.

2. PROCEDURE

The present study was based on a series of articles from Russian physics journals, comprising approximately 120,000 running words (some 500 pages). The processing of this text has been described elsewhere.[1,2] Here, we note only that each sentence of this text is recorded on magnetic tape, together with the following information for each occurrence in the sentence: its part of speech, its "word number" (an identification number in the machine glossary), and its syntactic "governor" or "dependent" (if any) in the sentence. A retrieval program applied to this text tape then yielded information about the SRW for words in which we were interested. For convenience and economy, all words in the machine printout for this study are identified by word number, rather than in their "natural-language" form.

In our study we chose to deal with the SRW of forty Russian nouns, herein called Test Words (TW). The number

is completely arbitrary; the particular nouns chosen (see Table 1) were presumed to form different semantic groupings. Table 1 gives one possible grouping of these words; the criteria for grouping are more or less obvious, although the reader may easily form different groups, by expanding or contracting the groups that we have designated. The only purpose of grouping is to provide a weak measure of control in the experiment: if two nouns are found to be similar in terms of their SRW, we should like to compare this finding with some intuitive understanding of their similarity. (For convenience, we shall refer to the TWs by their English equivalents.)

Two nouns may be compared with reference to several different types of SRW. Here, we have chosen to limit our comparison to three types: the adjective dependents (in either attributive or predicative function), the noun dependents (normally, but not necessarily, in the genitive case in Russian), and the noun governors (the TN is normally, but not necessarily, in the genitive case). Strictly speaking, the syntactic function of the SRW should be taken into account. In ignoring this factor, we are consciously permitting certain inexactitudes, on the premise that the distortions introduced into measurement will not be severe.

The task of manually retrieving SRW for each occurrence of the 40 TWs, and of comparing each TW with every other TW, is too tedious to be attempted. The aid of the computer was enlisted, in two ways.

Table 1

39 TEST NOUNS

		W No.	F	L1	L2	L3	L4
Group 1							
calculation[1]	vycislenie	782	62	15	23	11	49
measurement	izmerenie	1579	328	29	63	36	128
determination	opredelenie	3324	121	7	39	14	60
calculation[2]	rascet	4627	90	12	24	16	53
Group 2							
consideration	rassmotrenie	4598	51	14	29	6	49
comparison	sravnenie	5200	106	6	22	4	32
study	izučenie	1610	64	8	44	6	58
investigation	issledovanie	1723	159	32	65	21	118
Group 3							
relation	sootnošenie	5111	113	14	18	15	47
ratio	otnošenie	3455	102	14	22	9	45
correspondence	sootvetstvie	5109	29	2	1	0	3
Group 4							
solution	rastvor	4608	129	6	22	24	52
compound	soedinenie	5082	15	5	5	6	16
alloy	splav	5182	27	6	2	4	12
Group 5							
metal	metall	2460	86	11	2	28	41
gas	gaz	807	37	7	2	8	17
liquid	židkost'	1329	56	8	2	15	25
crystal	kristall	2131	171	15	19	44	78
Group 6							
uranium	uran	5745	171	0	0	18	18
silver	serebro	4899	48	4	1	17	22
copper	med'	2419	58	2	3	20	25
phosphor	fosfor	5913	130	9	2	34	45
Group 7							
proton	proton	4365	125	8	2	27	37
ion	ion	1686	98	14	10	31	55
molecule	molekula	2568	112	18	18	39	75
atom	atom	186	106	9	23	28	60
Group 8							
formula	formula	5911	231	20	21	19	60
expression	vyrazenie	739	223	25	12	24	61
equation	uravnenie	5742	412	42	24	32	98
Group 9							
width	sirina	6198	43	4	9	9	22
depth	glubina	913	40	6	8	9	23
length	dlina	1194	112	16	21	22	59
height	vysota	764	23	2	11	3	16
Group 10							
presence	nalicie	2696	119	3	73	5	81
absence	otsutstvie	3485	44	2	35	1	38
existence	sušcestvovanie	5352	41	3	25	6	34
Group 11							
question	vopros	613	96	5	3	10	18
problem[1]	zadaca	1362	68	15	11	10	36
problem[2]	problema	4254	26	4	10	6	20

"W No." = word number; "F" = frequency

1. Through automatic scanning of the text, each occurrence of the 40 TWs was located, and in each instance the identity (word number) of relevant SRW was recorded. A listing is produced for each of the TWs (see Table 2, "SRW Detail," for an example of the TW, VYČISLENIE = calculation[1]), showing the different words used as adjective dependents (List 1), noun dependents (List 2), and noun governors (List 3). The number of words on each of these lists is also shown in Table 1, together with the total number of SRW for each TW (List 4). We stress the fact that these numbers refer to <u>different</u> words used as SRW; the repetition of a given SRW (for a given SRW type) was not recorded.

2. Each TW was automatically compared with every other TW, with respect to their shared SRW, i.e., in terms of the words in Lists 1, 2, and 3 of the "SRW Detail Listing." A new listing, "Similarity Ranking by TW," is then produced (see Table 3 for the TW, VYČISLENIE = calculation[1]). This listing shows for each TW the number of shared SRW of each of the three types (N1, N2, and N3, Table 3), the total number of shared SRW (NA), and a measure of similarity for the pairs, herein designated as the Similarity Coefficient (SC). The SC is a decimal fraction obtained by dividing the sum of shared SRW for each pair of TWs by the product of the frequencies of the two TWs. (The latter is of course a device for taking into account the differing frequencies

Table 2

"SRW DETAILS"

LENGTHS OF LISTS, FREQUENCIES, AND RATIOS FOR TEST WORD NUMBERS

TEST WORD NUMBER = 782000 FREQUENCY = 62

LENGTH OF LIST 1 = 15 LENGTH / FREQUENCY = 0.24194

LIST 1 =

2004000 474000 1023000 6325000 6361000 4343000 994000 2768000 5440000 100000 2984000 2660000 2933000 5396000 3821000

TEST WORD NUMBER = 782000 FREQUENCY = 62

LENGTH OF LIST 2 = 23 LENGTH / FREQUENCY = 0.37097

LIST 2 =

1840000 4640000 1655000 5993000 4602000 4912000 9057000 6112000 2561000 453000 463000 2096000 1837000 4922000 5511000 5390000 2385000 1345000 1576000 6348000 5905000 3559000 2040000

TEST WORD NUMBER = 782000 FREQUENCY = 62

LENGTH OF LIST 3 = 11 LENGTH / FREQUENCY = 0.17742

LIST 3 =

2044000 9048000 4196000 585000 4674000 2466000 1528000 5763000 1362000 5390000 6364000

Table 3

"SIMILARITY RANKING BY TW"

ORDERED ON TWNI	DECIMAL TWNJ	JUDITENT NI	N2	N3	NA	(FI x FJ)	NA/(FI x FJ)
782000	2053000	0	0	0	0	62	0.
782000	5082000	0	0	0	0	930	0.
782000	5745000	0	0	0	0	1063	0.
782000	4899000	0	0	0	0	2976	0.
782000	2419000	0	0	0	0	3596	0.
782000	186000	0	0	0	0	6572	0.
782000	5198000	0	0	0	0	2666	0.
782000	913000	0	0	0	0	2480	0.
782000	764000	0	0	0	0	1426	0.
782000	613000	1	0	0	1	5952	0.00017
782000	5742010	5	0	1	6	25544	0.00023
782000	4608000	1	1	0	2	7998	0.00025
782000	4365000	2	0	0	2	7750	0.00026
782000	5911000	4	0	0	4	14322	0.00028
782000	2131000	2	1	0	3	10602	0.00028
782000	1329000	1	0	0	1	3472	0.00029
782000	2568000	2	0	0	2	6944	0.00029
782000	1194010	1	1	0	2	6944	0.00029
782000	1686000	2	0	0	2	6076	0.00033
782000	739000	4	1	0	5	13826	0.00036
782000	3485000	0	1	0	1	2728	0.00037
782000	2460000	2	0	0	2	5332	0.00038
782000	2696000	0	2	1	3	7378	0.00041
782000	807000	1	0	0	1	2294	0.00044
782000	5109000	1	0	0	1	1798	0.00056
782000	5182000	1	0	0	1	1674	0.00060
782000	5913000	4	0	1	5	8060	0.00062
782000	4254000	1	0	0	1	1612	0.00062
782000	1362000	2	0	1	3	4216	0.00071
782000	5200000	1	3	1	5	6572	0.00076
782000	5111000	3	3	0	6	7006	0.00086
782000	3455000	2	4	0	6	6324	0.00095
782000	1610000	1	1	2	4	3968	0.00101
782000	1579000	7	11	5	23	20336	0.00113
782000	5352000	0	1	2	3	2542	0.00118
782000	1723000	9	4	5	18	9858	0.00183
782000	3324000	1	9	5	15	7502	0.00200
782000	4598000	5	1	3	9	3162	0.00285
782000	4627000	7	6	5	18	5580	0.00323
0	0	0	0	0	0	0	0.

of the TWs; other means for determining this coefficient can be utilized.) The pairings for each TW are ordered on the value of the SC. It should be noted that the similarity between TWs is measured in terms of the total number of shared SRW (Column NA of Table 3); it is also possible to express this measurement in terms of shared SRW of any single type.

A third listing was also produced: a listing of the 780 TW-pairs, ordered on the value of the SC. This listing, not reproduced here because of its length, will be referred to as "Ranking of TW-Pairs by SC." Table 4 shows the distribution of the SC as compared with the number of TW pairs.

The following discussion is based on the three listings described above. A few additional remarks may be made about the procedure itself, which may be likened to deep-sea fishing with a tea strainer full of holes. The limitations of size are obvious: we have limited ourselves to three of the numerous ways of comparing nouns in terms of their SRW. Other types of SRW that suggest themselves are: verbs, where TW is subject; verbs, where TW is direct object; prepositional phrases as dependents, or governors, of TW; nouns joined to TW through coordinate conjunctions (i.e., "apples" and "grapes" are said to be more similar if "apples and oranges" and "grapes and oranges" occur in text). Some of the holes in our tea strainer are: the neglect of the case of the noun dependent of TW, or the

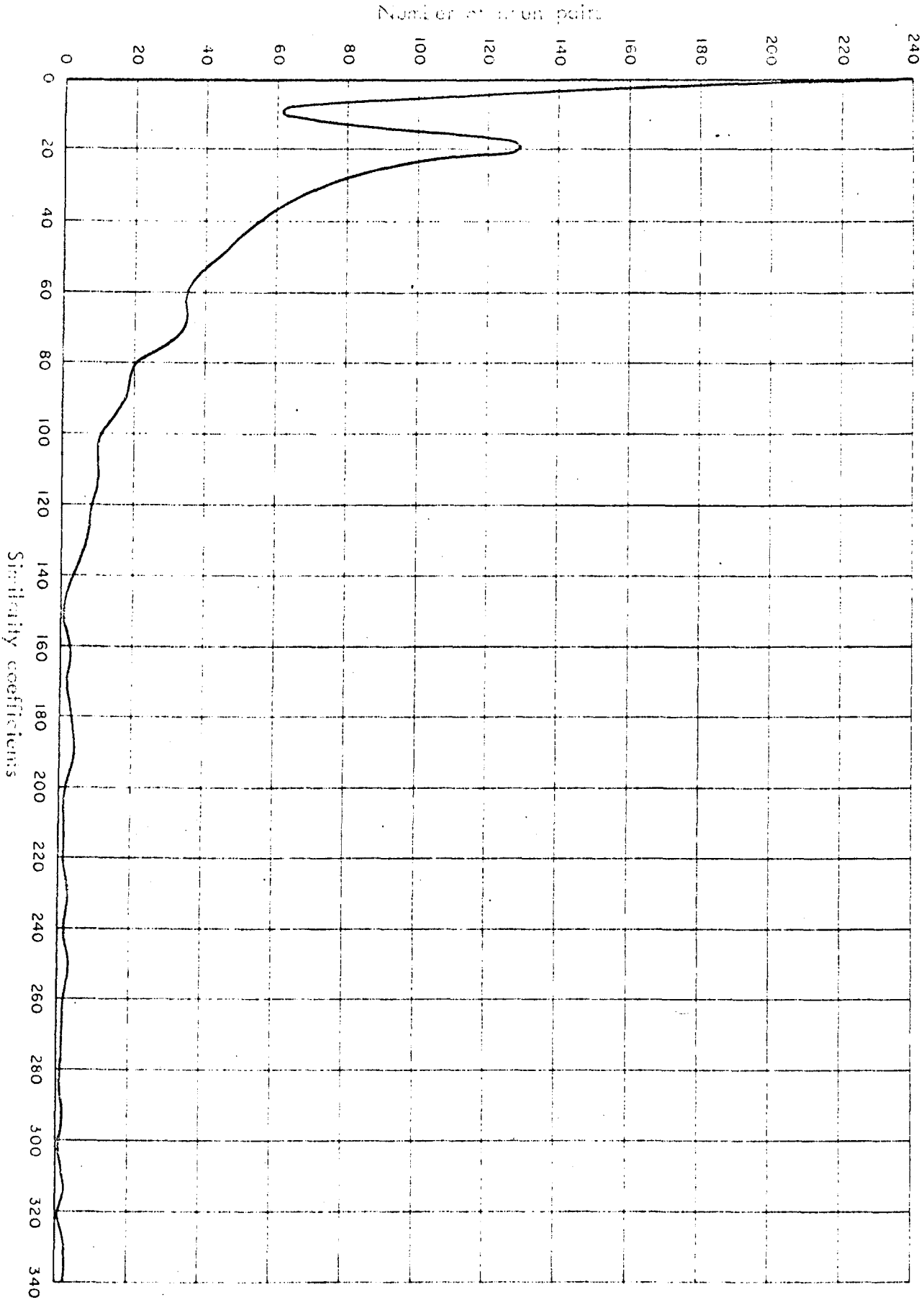

Table 4—Distribution of similarity coefficients

case of the TW when the SRW is a noun governor; the neglect of technical symbols in physical text, as dependent or governor of the TW; the failure to distinguish between different functions of governors or dependents in a noun/noun pair (e.g., the distinction between "subjective" and "objective" genitive); the neglect of transformationally equivalent constructions. In view of these deficiencies (not to mention the problem of statistics), the success of our fishing expedition is open to doubt. Let us then proceed to examine the catch.

3. RESULTS

The evaluation of the data contained in our three machine listings is not an easy task. We can scarcely examine and discuss the degrees of similarity of 780 noun-pairs. The problem of interpretation is also complicated: how completely and accurately should the results correspond with our expectations, as represented in the tentative semantic groupings (Table 1)? Our approach is to deal in a summary manner with the noun-pairs characterized by highest Similarity Coefficients, especially with respect to their intra- and inter-group relationships. Before proceeding to this discussion, a few preliminary remarks should be made about the data in the various machine listings.

The summary of SRW counts for each TW, contained in Table 1, suggests all TWs do not have the same opportunity for comparison. In the case of "correspondence" (Group 3),

a total of only three SRW is noted in (Column 14); as a result, this TW should be eliminated from further consideration. In addition, unless at least two, and preferably all three, types of SRW are well represented for a given TW, the SC for that noun will tend to be skewed. As examples, we note all nouns in Group 6 (for which the L3 column predominates), and the nouns in Group 10 (for which the L2 column predominates). In effect, these nouns are "deficient" in certain types of SRW, and require special handling.*

On the printout, "Ranking of TW-Pairs by SC," a number of noun pairs appear at the top end of the scale although the total number of shared SRW is small (i.e., the value of column "NA" (see Table 4) is "1," "2," or "3." The SC may be high, because the product of the frequencies is relatively low. Our policy has been to discount these pairs on the grounds that the value of "NA" is significant in determining the similarity between two TWs. The minimum value for NA was arbitrarily set at four.

Keeping in mind these amendations to the data in mind, we proceed to the discussion of the noun-pairs characterized by highest SC. Table 3 shows the distribution of SC by noun-pairs. By any standard, the data shows negative or extremely weak similarity for most of the 780 pairs.

---

*An abstract of a paper on the proclivity of nouns to enter into certain combinations is cited in Reference 3.

At which point on the curve shall we draw a line, saying that an SC above this value indicates similarity, and that an SC below this value indicates dissimilarity or weak similarity (all this of course in terms of reliability)? For purposes of discussion, we propose to set the threshold at .00100--a rigorously high figure. After eliminating pairs whose NA value is less than 4, we find 38 pairs whose SC lies in the range .00100 to .00337 (Table 5). (The first two zeroes are dropped.)

The reader may draw his own conclusions about the degree of similarity between the nouns in any given pairing. For purposes of discussion, we will refer to the pairings in terms of our preliminary groupings (Table 1). The following intra- and inter-Group pairings are observed in Table 5:

Nouns of Group	pair with nouns of Group
1	1, 2
2	1, 2, 10
3	-
4	5
5	4, 5, 6, 7
6	5, 6, 7
7	5, 7
8	-
9	9
10	2, 10
11	5, 11

We note that no pairings appear for nouns of Groups 3 and 8. All other groups except Group 4 are represented by intra-group pairings; to this degree, our expectations are fulfilled, i.e., the data supports our a priori feelings for the similarity between words. The amount of inter-

Table 5

"HIGH RANKING TW-PAIRS"

TWI	Group	TWJ	Group	SC	NA
calculation[1]	1	calculation[2]	1	323	18
		consideration	2	285	9
		determination	1	200	15
		investigation	2	183	18
		measurement	1	113	23
		study	2	101	4
determination	1	calculation[2]	1	165	18
study	2	consideration	2	337	11
		existence	10	267	7
		investigation	2	246	25
		absence	10	213	
		calculation[2]	1	139	8
		presence	10	118	9
		determination	1	116	9
investigation	2	consideration	2	173	14
		calculation[2]	1	154	22
		absence	10	114	8
		existence	10	107	7
consideration	2	calculation[2]	1	174	8
liquid	5	molecule	7	143	9
		problem[1]	11	105	4
		metal	5	125	6
		crystal	5	104	10
gas	5	metal	5	126	4
metal	5	silver	7	194	8
crystal	5	compound	4	156	4
copper	6	silver	6	180	5
		metal	5	120	
ion	7	copper	6	106	6
atom	7	ion	7	125	13
height	9	length	9	155	4
depth	9	width	9	233	4
length	9	width	9	125	6
absence	10	existence	10	222	4
		calculation[2]	1	101	4
presence	10	absence	10	229	12
		existence	10	225	11
question	11	problem[2]	11	240	6

group pairing may indicate either that the data is inconclusive, or that our original groupings were too narrow. In fact, two larger groups emerge: one composed of Groups 1 and 2 (perhaps including Group 10), the other composed of Groups 4, 5, 6, and 7. This tendency is more marked if we lower the SC threshold from .00100 to .00070, thereby adding a total of 28 pairs to the number listed in Table 5. For example, nouns of Group 1 are found to pair with those of Group 10, and nouns of Group 4 pair with those of Groups 6 and 7.

The data is not statistically conclusive, but strongly suggests the emergence of the two major groups mentioned above. The amalgamation of Groups 1 and 2 can easily be defended on semantic grounds; since Group 10, as noted above, is subject to aberrant behavior (because of the very high number of noun dependents), its inter-relation with Groups 1 and 2 may not be taken seriously.* Groups 4, 5, 6, and 7, which include the names of chemical mixtures, classes of elements, individual elements, and components of elements, may be taken together semantically as a single sub-class of "object nouns." The physicist tends to say the same things about all nouns in this group.

One of the 38 pairs listed in Table 5 appears to contradict expectation: "liquid"/"problem"(Groups 5 and 11).

---

*It should also be noted that the noun dependents of Group 10 nouns serve a "subjective" rather than "objective" function. If we had distinguished between the syntactic function of the noun dependent, TWs of Group 10 would be only weakly similar to TWs of Groups 1 and 2.

The four SRW shared by these two nouns include the adjective "certain" and the noun governor "number." The non-discriminatory ("promiscuous") nature of these two SRW is perhaps obvious, and one of the refinements that should be introduced in future studies is the neglect of such words as "significant" SRW. (The study of "promiscuity" in adjectives is referred to in Reference 4.) At the present, experience suggests that distortions introduced by such words are minimal if the number of SRW is sufficiently large.

Our general conclusion is that, with a few anomalies, the 66 pairings for which the SC is .00700 or higher meet with our expectations.

Another aspect of the question remains: many nouns with presumed similarity are not represented on the high end of the SC distribution curve. (If we lower the threshold to include such pairs we shall also encounter many non-similar pairs.) One way of dealing with this problem is to consider the most highly correlated pairs that nouns in each Group form, whether or not the SC is "significantly" high. In lieu of presenting this information in full detail, we show in Table 6 the most closely correlated pairs for a representative noun from each of the Groups (excepting Groups 3, 4, and 8).

The most striking aspect of Table 6 is the repetition of intra- and inter-Group pairings noted in Table 5 for high-SC pairings. In other words, the relative value of

Table 6

PAIRS WITH HIGHEST CORRELATION

(one example from each Group)

	Pairing	SC	Grp.
Group 1 (calculation[1])	calculation[2]	323	1
	consideration	285	2
	determination	200	1
	investigation	183	2
	measurement	113	1
	study	101	2
	ratio	95	3
	relation	86	3
	comparison	76	2
Group 2 (consideration)	study	337	2
	calculation[1]	285	1
	calculation[2]	174	1
	investigation	173	2
	determination	97	1
	comparison	92	2
	measurement	60	1
Group 5 (metal)	silver	194	6
	gas	126	5
	liquid	125	5
	copper	120	6
	crystal	68	5
	phosphor	63	6
	atom	55	7
	molecule	52	7
	ion	47	7
	solution	45	4
	proton	37	7
Group 6 (uranium)	silver	49	6
	copper	40	6
	ion	18	7
	atom	17	7
	gas	16	5
	molecule	16	7
	metal	14	5
	liquid	10	5
Group 7 (proton)	ion	90	7
	atom	60	7
	gas	43	5
	copper	41	6
	metal	37	5
	phosphor	37	6
	molecule	36	7
	crystal	33	5
Group 9 (width)	height	303	9
	depth	233	9
	length	125	9
	ratio	68	3
Group 10 (presence)	absence	229	10
	existence	225	10
	study	118	2
	consideration	82	2
	investigation	74	2
Group 11 (question)	problem[2]	240	11
	problem[1]	46	11
	formula	18	8
	equation	15	8

the SC appears to be as significant as the absolute value. This result was certainly not expected, and perhaps indicates a greater sensitivity in our measurement procedures than we would have thought reasonable.

Table 6 suggests, but does not prove, the existence of clusters (or "clumps") of TWs, in which the members are closely correlated with each other, and in which no member is closely correlated to any outside word. We have not yet attempted to apply clumping procedures; a better understanding of the data is perhaps a prerequisite to this rigorous treatment. For the present, we shall point out a phenomenon that strongly suggests the existence of clumps: the recurrence of the same SRW among several TWs with high mutual correlation. Consider, for example, that a high SC is found between Test Words A and B, B and C, and A and C; if, in addition, a relatively high proportion of SRW are shared by all three TWs, the mutual connection of the three words would appear to be considerably strengthened. The recurrence of SRW has not been systematically studied, but the following sample is offered as an illustration of the phenomenon. Below, we list all the SRW of the three types, for the TW calculation[1]. The underlined words are those which, in addition, also served as corresponding SRW for two other TWs (determination, and measurement) that are highly correlated to each other and to calculation[1].

Table 7

SRW OF CALCULATION[1]

Adjective Dependents: (L1)	TAKOJ (such); ANALOGICNYJ (analogous); DAL'NEJSIJ (further); NAŠ (our); NEPOSREDSTVENNYJ (direct).
Noun Dependents: (L2)	ZAVISIMOST' (dependence); MASSA (mass); VELIČINA (magnitude); SECENIE (cross-section); KOEFFICIENT (coefficient); MODUL' (modulus); RASSTOJANIE (distance); SILA (force); FORMA (form).
Noun Governors: (L3)	ZRENIE (view); REZUL'TAT (result); VOZMOŽNOST' (possibility); METOD (method).

Table 7 shows that eighteen SRW appeared for calculation[1]. Of these, one half (nine) also appeared as SRW for both determination and measurement. It would seem that the "togetherness" of these three TWs is strengthened by this feature, which we term "recurrence of SRW." We have no ready formula for determining that recurrence is or is not significant in a given situation. In general, the nature and behavior of individual SRW remain to be studied, so far as their relevance to our problem is concerned.

4. CONCLUSIONS

We conclude that there is considerable agreement between the results of our experiment and an a priori feeling for the similarity of words. Words that are similar in meaning tend to have the same SRW, to a far greater degree than chance would determine. If this conclusion is valid, a large-scale experiment is suggested, using a larger number of Test Words, more SRW types, and a larger

amount of text. (The text base for the present experiment proved to be adequate; larger amounts of text should, however, remove some of the anomalies.) The question of further refinements in the procedure must also be taken seriously: e.g., we may also take into account multiple occurrences of an SRW, distinguish to some degree the different functions of noun governors or noun dependents, discount the occurrence of "promiscuous" SRW. Clumping procedures should be applied, perhaps taking into account the recurrence of individual SRW among a group of Test Words.

REFERENCES

1. Hays, D. G., and T. W. Ziehe, Russian Sentence-Structure Determination, The RAND Corporation, RM-2538, April 1960.

2. Hays, D. G., Basic Principles and Technical Variations in Sentence-Structure Determination, The RAND Corporation, P-1981, April 1960.

3. Harper, K. E., "A Study of the Combinatorial Properties of Russian Nouns," Mechanical Translation, August 1963, p. 36.

4. Harper, K. E., Procedures for the Determination of Distributional Classes, The RAND Corporation, RM-2713, January 1961.