

MEANING, WORD COMBINATION AND STATISTICS

Direct data on language units and their properties may be obtained by analyzing them in written texts and speech. Taking into account the ever increasing role of written texts in every-day life and keeping in mind that they are more accessible for experimental purposes than speech, we found it preferable to use them for studying the semantics of lexical units.

A word (*lexeme*) and its *meaning* are language phenomena, that is why we perceive words indirectly through their usage in speech/texts. In written texts words are realized in the form of specific, visually recognizable *word usages* (i.e. chains of letters between two blanks in a text line) and *word forms* (i.e. externally similar word usages possessing identical grammatical and semantic properties). For example, the word *срón* (table) in Russian is represented by twelve word forms (six cases in the singular and six cases in the plural) and, for a given text, a certain number of word usages. Hence, a word as language unit is manifested in a text through the whole combination of word forms and word usages from which, in turn, this word is abstracted.

Since semantic properties of a word are realized only in its usage¹, and different meanings of a word in the long run are manifested by different combinations with other words, it is but natural to analyze the words' semantic properties by observing its combinatorics.

According to Ch. Peirce², one of the founders of semantics, it is accepted practice in modern linguistics to distinguish several aspects of word semantics. The relation of the word's material envelope to an object in the real world is called the denotative aspect of the word's meaning; the relation of the word's material envelope to the notion, triggered by it in our mind, is called the significative aspect of the word's meaning; the relation of the word to the person using it is

called the pragmatic aspect of the word's meaning. Apart from these semantic aspects, F. de Saussure has delineated the linguistic aspect of meaning (value), which comprises the whole complex of the word's relations to other words in the language³. In analyzing word combinations we try to describe the linguistic aspect of meaning, because similarity and differences of semantic properties are more clearly manifested in similarity and differences of word combinations. Yet it should be noted that words are combined with respect to all the above mentioned aspects of their meaning, the primary importance being attached to the word's correlation with objects and realities, i.e. the denotative aspect of semantics. In practice one does not meet a word combination like *деревянное железо* (wooden iron) or *кирпич улыбается* (the brick is smiling). This is because such word combinations do not correspond to the real world; on the other hand, word combinations like *солнце светит* (the sun is shining), *голубое небо* (blue sky) and so on are frequently met in texts and sound quite normal to us, since the lexemes of each such combination denote definite realities.

It should be noted that in normal discourse (texts) often and easily only those words are combined which have meaning elements or specific semantic signs in common. For example, the words *солнце* (the sun) and *светит* (shines) do have such a common semantic sign, so it is but natural that they will combine one with another. Furthermore, words normally possess many specific semantic signs which are revealed in combinations of these words with others. If one considers word combinations as a function of the semantic mechanism of language, i.e. as a specific realization of the semantic properties of the words being combined, it may be easily concluded that a polysemantic word (a word having a great number of semantic signs) should combine with a greater variety of lexemes than a word possessing fewer semantic signs, or a term.

To justify this statement, the following statistical linguistic experiment has been carried out. From many texts representing different styles of the Russian language were

extracted, by thorough selection, word combinations with the words земля (earth), солнце (sun), небо (sky), звезда (star), луна (moon), сердце (heart), рука (hand), нога (leg), глаз (eye). For each lexeme, one thousand two-word combinations have been selected from scientific texts, poetry and literary prose, respectively, that is 27 thousand word combinations in all, comprising three thousand for each above-mentioned word. Since in scientific texts words are used as terms and in literary texts they are polysemantic, then, according to our hypothesis, the scope of word combinations in scientific texts should be considerably smaller than in literary texts. After extracting all the relevant word combinations from the texts under study and counting the different partner words for each thousand word usages, we received the following results (See Table 1).

Table 1

Results received Lexemes	Number of different partner words per each thousand of word combinations with given lexemes in our texts		
	scientific texts	literary prose	poetry
звезда (star)	229	506	608
луна (moon)	297	494	594
небо (sky)	277	534	627
земля (earth)	275	610	700
солнце (sun)	268	478	605
глаз (eye)	286	540	599
рука (hand)	229	564	610
сердце (heart)	195	576	653
нога (leg)	193	585	643

The data reveal a clear-cut tendency: a sharp increase in collocational diversity in fiction (prose and poetry) as compared to scientific texts. Hence, our assumption appears to have been borne out by the results of the experiment: polysemantic lexemes have a greater scope of combinations with

different words than lexemes used as terms, i.e. monosemantic words.

It should be underlined that the semantic signs of a lexeme, expressed in its combinations, may be *specific* as well as *general*. We find specific signs in each new combination with another word; general signs are found in whole groups of semantically similar words which combine with the given lexeme. For example, the specific semantic sign of colour, pertaining to the object denoted by the word дерево (tree), may be expressed separately by adjectives combining with the word дерево (tree), such as: зеленое (green), белое (white), желтое (yellow), красное (red) and so on. The general semantic sign of the word дерево, the sign of colour, is expressed by all adjectives denoting colour which combine with it. Another general semantic sign of the word дерево may be referred to by adjectives of size like большое (big), маленькое (small), огромное (huge), высокое (tall), and so on.

By studying a great number of texts comprising millions of word usages, one may find dozens of such general semantic signs which are expressed in semantically similar words combining with the lexemes under study. It should be stressed that different words, on the one hand, may not have equal sets of semantic signs (that depends upon the grammatical and semantic properties of the words); on the other hand, the *size* (volume, weight) of semantic signs (defined by the frequency of word usages in relevant combinations) may also not correspond. So, words having similar meanings (synonyms) apparently differ from one another not by the presence/absence of certain semantic signs in their semantic structure, but by the sizes of these signs. Besides, the meaning of each word comprises not a mere set of signs, but the signs specifically ordered in a word structure⁴. This order of semantic signs is manifested in the way the word combinations with the given lexeme are realized in specific syntactic constructions.

For example, the semantic sign 'feeling' peculiar to the meaning of the word человек (man) may be manifested in verbs like любить (love), ненавидеть (hate), обожать (worship), презирать (despise) and so on, which are all capable of combining with the noun человек: but this sign may differently

characterize the meaning of человек, depending on the syntactic construction in question. If the verbs of feeling are used in syntactic constructions of the type "given noun in the Nominative case + verb" (compare: человек любит (the man loves), человек ненавидит (the man hates, and so on), then the semantic sign manifested by these verbs is active. On the other hand, the sign is considered passive when the verbs of feeling are used in syntactic constructions of the type "verb + given noun in the Accusative case" (compare: ненавидеть человека (to hate the man), любить человека (to love the man, and so on). In general, semantic signs, when realized in word combinations of different syntactical structure, are considered different despite the fact, that they may be expressed by the same lexemes.

Hence, lexical meanings are defined by generalized semantic signs which are peculiar to whole groups of words combining with the given lexemes. The statistical analysis of word usage in each group makes it possible "to weigh" the semantic signs' quota in the given lexeme structure. The semantic structure of the lexeme (i.e. the order of its semantic signs) is manifested by the distribution of lexical groups expressing generalized semantic signs, and by the syntactic constructions peculiar to the given lexeme.

According to this method, proposed by A.E. Suprun, we have carried out a statistical linguistic experiment with the following eight words: мысль (thought), идея (idea), понятие (notion), мнение (opinion), ум (intellect), разум (intelligence), сознание (mind), мышление (thinking), which, to our mind, are quite related in meaning. Our task has been to define the degree of similarity in the semantic structure of these words by analyzing empirical data on their collocational properties. All word combinations with the given lexemes were extracted by thorough selection from Russian texts containing more than two million word usages. There were found: 3210 combinations with the word мысль (thought), 2087 - with the word идея (idea), 1245 - with the word понятие (notion), 881 - with the word ум (intellect), 836 - with the word сознание (mind), 628 -

with the word мнения (opinion), 484 - with the word разум (intelligence), 416 - with the word мышление (thinking). First of all, the word usages combining with the lexemes under study were arranged into syntactic constructions of the following types: мысль писателя (the thought of the writer), суть мысли (the essence of the thought), высказать мысль (to express the thought), etc. - all in all 10 syntactic constructions peculiar to these lexemes were established. In the following stage, within these syntactic categories word usages were arranged into lexical groups, each representing a separate semantic sign of the given eight lexemes. We devised the so-called "semantic questionnaire", the "questions" relating to the names of the semantic signs peculiar to the meaning of our eight words, and the "answers" being quantitative dimensions of each sign (the dimensions are given by the frequency of word usages having this or that semantic sign).

The group of lexemes under consideration, with the given syntactic constructions, is characterized by 72 semantic signs each of them being represented by a lexical group with a certain number of word usages (from 2 to 1218). The frequency of word usages taken from our texts expresses quantitatively the dimensions of the corresponding semantic signs.

However, since word frequencies differ with texts, the dimensions of semantic signs should be expressed not only by the definite frequencies of word usages which fall into the specific lexical groups but also by the proportion of frequencies in per cent in relation to the general frequency of all word usages with each given lexeme.

Table 2 shows the 20 most significant (by size) semantic signs characterizing the meaning of the words under study; the word usages of each group constitute not less than 2 per cent of all word usages extracted from our texts; each of the remaining 52 signs, not shown in Table 2, constitute a proportion of word usages equal to 1 per cent or even less.

By comparing the data of Table 2, we get information about similarity or difference in the sizes of corresponding semantic signs of the lexemes under study.

Table 2

Syntactic construction	Lexemes Lexical group								
		мысль	идея	мнение	понятие	ум	разум	сознание	мышление
1	Pronominal adjectives (that, my etc.)	468 15	241 12	115 18	162 13	112 13	39 9	59 7	22 5
	Intellect (clever, sensible etc.)	102 3	83 4	2 0	47 4	22 2	7 1	12 1	31 8
	Totality (group, collective etc.)	123 4	53 3	60 10	18 1	9 1	9 2	42 5	3 1
	Distinctiveness (obvious, visible etc.)	83 3	89 4	9 2	54 4	12 1	12 2	9 1	9 2
	Qualitative (good, bad etc.)	298 9	155 8	40 6	85 7	74 8	22 5	33 4	23 6
	Size (big, small etc.)	79 3	71 3	16 3	33 3	69 8	15 3	26 3	5 1
2	Names (man, writer, boy etc.)	112 4	36 2	38 6	15 1	57 8	28 6	68 8	13 3
	Proper names (Peter, Pushkin etc.)	177 6	130 6	97 16	11 1	40 5	6 1	27 3	11 3
	Intelligence (conception, theory etc.)	28 1	86 4	3 0	52 4	3 0	1 0	10 1	10 2
3	Quality (fidelity, narrowness etc.)	69 2	18 1	5 1	20 2	38 4	16 3	11 1	21 5
	Process (development, movement etc.)	139 4	96 5	18 3	48 4	20 2	27 6	29 3	9 2
	Intelligence	79 3	46 2	3 0	32 3	7 1	31 7	9 1	37 9
4	Motion (to go, to run etc.)	122 4	53 3	9 2	20 2	24 3	13 2	10 1	12 3
	Existence (to be, to live etc.)	131 4	95 5	19 3	53 4	23 3	26 6	29 3	24 6
	Intelligence (to think, to decide etc.)	57 2	51 2	11 2	31 3	28 3	17 4	22 3	10 2
5	Intelligence (to express etc.)	179 6	90 4	31 5	28 2	12 1	9 2	14 2	7 2
	Displacement (to pull, to carry etc.)	95 3	47 2	12 2	56 5	19 2	10 2	23 3	5 1
6	Intelligence (to seek etc.)	33 1	40 2	15 2	16 1	15 2	9 2	21 3	10 2

Table 2 continued

6	Motion	27 1	29 1	6 1	19 2	27 3	1 0	44 5	6 1
	Displacement	34 1	24 1	7 1	34 3	8 1	8 2	32 4	4 1

Note. In the left column, the numbers from 1-6 stand for the following types of syntactic construction: 1. adjective + given noun (good idea), 2. given noun + other noun dependent on it (thought of the writer), 3. other noun + given noun in the Genitive case (narrowness of the thought), 4. given noun in the Nominative case + Verb (the idea lives), 5. verb + given noun in the Accusative case (to express the thought), 6. verb with preposition + given noun in an oblique case (to seek for the idea). In these syntactic constructions we met the most significant (in terms of frequency) groups of word usages in combination with the given nouns. Opposite each group of word usages there are two rows of numbers: the upper row shows the frequency of word usages of the given group, the lower row shows the volume of these numbers in proportion to the sum of word usages with the given noun expressed as a percentage.

To measure the degree of similarity in the meanings of words by semantic signs, we use the statistical formula of the correlation ratio⁵:

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

where, r-correlation ratio (relation measure), x_i and y_i - the definite sizes of the semantic signs in pairs of compared words, $\bar{x}$ and $\bar{y}$ - the average sizes of semantic signs. With the help of the correlation ratio the relation measure is expressed within the limits of -1 to +1. The closest measure of direct relation of two objects is +1. Hence, a result nearer to +1 testifies closer direct relation of the two objects compared.

By calculating the relation ratio between combinatory properties of each pair of words in our group, we got the results shown in Table 3.

Table 3

	мысль	идея	мнение	понятие	ум	разум	сознание	мышление
мысль (thought)		0,93	0,85	0,84	0,84	0,71	0,64	0,61
идея (idea)			0,80	0,86	0,77	0,65	0,61	0,62
мнение (opinion)				0,62	0,72	0,51	0,65	0,36
понятие (notion)					0,72	0,61	0,60	0,57
ум (intellect)						0,74	0,72	0,57
разум (intelligence)							0,60	0,74
сознание (mind)								0,43
мышление (thinking)								

The highest relation incidence (by combination of all semantic properties realized by our eight words in texts) was found for the lexemes *мысль* (thought) and *идея* (idea) - 0,93, *идея* (idea) and *понятие* (notion) - 0,86; the weakest relation incidence was determined for the words *мнение* (opinion) and *мышление* (thinking) - 0,36, *сознание* (mind) and *мышление* (thinking) - 0,43.

These results may be used for interpreting such semantic phenomena as synonymy: words with high relation incidence (correlation ratio of the order 0,7 or higher) may be considered as strong synonyms, while words with low relation incidence may be considered as weak synonyms. In our analysis words are classified not on the principle "synonym/not synonym", but rather by the degree of synonymical relations determined between these words, i.e. on the basis of experimental analysis of their combinations⁶.

С Н О С К И

1. А.А. ПОТЕВНЯ, *Из записок по русской грамматике*. Т. 1-2. М., 1958, с. 42. Л. ВИТГЕНШТЕЙН, *Логико-философский трактат*. М., 1958, с. 39.
2. *The Philosophy of Peirce. Selected Writings*. New York, 1940, p. 101.
3. Ф. де СОССЮР, *Труды по языкознанию*. М., 1977, с. 142.
4. F. HUNDSNURSCHER, *Neuere Methoden der Semantik*. Tübingen, 1971, S. 19. Ю.Д. АПРЕСЯН, *Лексическая семантика*. М., 1974, с. 9.
5. Л.С. КАМИНСКИЙ, *Измерение связи (корреляция)*. Л., 1962.
6. The present article is based on a former book of the author where the interested reader can find further results concerning lexical meaning: В.А. ПЛОТНИКОВ, *Дистрибутивно-статистический анализ лексических значений*. Минск, 1979.
Other works in the field are the following:
А.Е. СУПРУН, В.А. ПЛОТНИКОВ, Манифестация значения и его исследование. - В кн.: *Проблемы семантики*. М., 1974, с. 255-262.
В.А. ПЛОТНИКОВ, Об использовании лексикографических данных при построении тезауруса. - *Научно-техническая информация*, сер. 2, №9, 1975, с. 22-23.
В.А. ПЛОТНИКОВ, Лексическая сочетаемость и значение слова в тексте. - В кн.: *Восприятие языкового значения*. Калининград, 1980, с. 64-71.
В.А. ПЛОТНИКОВ, О типологии связей между формой и значением слов. - *The Prague Bulletin of Mathematical Linguistics*, 42, 1984, p. 27-36.
В.А. ПЛОТНИКОВ, Сочетаемость и семантика. - В кн.: *Методы изучения лексики*. Минск, 1985, с. 74-81.