Barbara Starosta

THE NOTION OF INFORMATION AND TEXT ANALYSIS

Originally published as ”Pojęcie informacji a analiza tekstu,” Studia Semiotyczne 7 (1977), 127–143. Translated by Agnieszka Ostaszewska.

1. INTRODUCTION

In semiotic works one often employs the notion of information. Authors use it mainly in descriptions of the communicational role of language signs. Many works are devoted to the problems of transfer, recording, processing and searching for information. A separate issue is in obtaining information on the cognitive process and the necessity to use the term ”information” in works on cognitive tolls, which include prepositions, theories or definitions. Connected therewith is research concerning the inference methods and in particular non-deductive inference methods.

The information theory, through the attainments of recent years, is not sufficiently spread in the circles of semioticians, and therefore the notion of information is usually employed in accordance with the intuitive feeling of a given author. In the situation, where on one hand there exists a clear need to use the notion of information in semiotic works, and on the other hand, where the theory of information disposes of an explication of the notion of information, it seems interesting to try the usefulness of the information theory for the purposes of language research.

The term ”information” similarly to ”probability” belongs presently to well-defined expressions of the scientific dictionary and is convenient to use in those cases when we are dealing with events. I will not discuss the notion of an ”event,” I assume that the sense thereof is known. Generally, we call an event results of experiments or observations. These results are recorded in a certain language. E.g. a connection obtained as a result of the dialling of a certain number on the telephone is an event, whose record is the telephone number indicated in the phone book. The fact that Johnny is going to school is an event, which is recorded as ”Johnny is going to school.”

I limit my discussions only to an analysis of records, and more strictly speaking to an analysis of written texts in the Polish language. The notions of language and text which I will be using, are defined in more detail in part 4.

In this paper I use the notion of information characterised by J. Kampé de Fériet in the publication *La Théorie Généralisée del Information et la Mesure Subjective de l'Information* (1974) and moreover I use the axiomatics provided by J. Losfeld in *Information Généralisée et Relation d'Ordre* (1974). Kampé de Fériet's article is accompanied by a rich bibliography, which lists almost all the most important works on information theory published in the years 1967-1974. The books used by me for discussion of the topic, have been listed in the footnotes and in the bibliography.

2. THE NOTION OF INFORMATION

Most generally, we define information as a function defined on a class of subsets of a non-empty set Ω, whose main property is monotony for inclusion. This means that: if with an S we mark the class of subsets Ω, with $\leqslant$ we mark the relation of partial order (reflexive, transitive, asymmetrical) on set S, and with R^+ we mark the set of positive real numbers, and with function I, the function of information, then we will obtain the following notation:

1.1. I: $S \rightarrow R^+$

1.2. For each A, $B \in S$

$A \subseteq B \Rightarrow \mathrm{I}(A) \geqslant \mathrm{I}(B)$

The above definition of information does not impose any conditions upon sets Ω and S, apart from the fact that class S is a set partially ordered by the relation of inclusion. For example, we interpret set S as a set of sentences ordered by the relation of logical resulting, then according to 1.2., if from sentence p there follows logically sentence q, then information of sentence p is greater than the information of sentence q. Which we may write down as follows:

1.3. $p \rightarrow q \Rightarrow \mathrm{I}(p) \geqslant \mathrm{I}(q)$

If for the relation of order in S there exists the smallest element m and the biggest element M, then we assume that:

1.4. $\mathrm{I}(m) = +\infty$
1.5. $\mathrm{I}(M) = 0$

If empty set $\varnothing$ and set Ω belong to S, then the smallest element of set S is the empty set, since it is contained in any set, and the largest element of set S is the full set Ω. According to 1.4. and 1.5., information of the empty set is equal to $+\infty$, and the information of the full set Ω is equal to 0.

1.6. $\mathrm{I}(\varnothing) = +\infty$
1.7. $\mathrm{I}(\Omega) = 0$

The above establishments agree with our intuition connected with the notion of information. If, for example, John and Mary informed us that they would come to us tonight for dinner, then the repeated message that they give to us does not provide any new information. On the other hand, a message that Mary and John will bring their pet crocodile to the dinner might be surprising for us. The measure of information is therefore in this case the degree of our surprise, astonishment.

When formulated differently, the same thought is expressed in the following manner: we obtain information always and only as an answer to a question, in other words, information is a function of a question. If in the answer we obtained only what we already knew, then the answer does not provide any information, it is not simply an answer to the question asked. An answer brings more information the more it liquidates the degree of our ignorance.

Going back to the example with a set of sentences, an always false sentence $p \wedge \sim p$ is the smallest element of the set of sentences S, since from this sentence every other sentence results. An always false sentence is indefinitely informative:

1.8. $\mathrm{I}(p \wedge \sim p) = +\infty$

An always true sentence ($p \cup \sim p$) has, by such interpretation, zero information. If we mark an always true sentence with t, then the above statement shall be noted in the following manner:

1.9. $\mathrm{I}(t) = 0$

Set S, partially ordered by the relation of order, having the smallest and the largest element and closed to the set theory operation of union and multiplication, shall be called a check. In case of such an algebraic system, for each A, $B \in S$, $A \cap B$is an element of S and $A \cup B$ is an element of S. If set S, on which we determine the function of information, is a check, then we may introduce the notion of independence of set, with the use of the notion of information in the following manner:

1.10. A independent of $B \Leftrightarrow \mathrm{I}(A \cap B) = \mathrm{I}(A) + \mathrm{I}(B)$.

Two sets, A and B, A, $B \in S$, are independent if and only if the information of their product is equal to the information contained in each of those sets. For example, if set S of sentences is closed due to conjunction $\wedge$ and alternative $\vee$, then sentence p is independent of sentence q if and only if the information contained in the conjunction of these two sentences is equal to the sum of information in these two sentences:

1.11. p is independent of $q \Leftrightarrow \mathrm{I}(p \wedge q) = \mathrm{I}(p) + \mathrm{I}(q)$.

Formula 1.11. reflects our intuitions well: information of the sentence ”Johnny is ill and primroses have blossomed today” is equal to the information of the sentence ”Johnny is ill” plus the information of the sentence ”Primroses have blossomed today.” The sentences ”Johnny is ill” and ”Primroses have blossomed today” are independent of each other.

If set S has the structure of a check, then on this set we define conditional information provided by element A of subset S by knowledge of element B of subset S. For each A, $B \in S$

1.12. $\mathrm{I}(A/B) = \mathrm{I}(A \cap B) — \mathrm{I}(B)$ if $\mathrm{I}(B) < +\infty$
1.13. $\mathrm{I}(A/B) = \mathrm{I}(A)$ if $\mathrm{I}(B) = +\infty$

For example, information in the sentence ”Johnny has a flu, provided that Johnny is ill,” if the sentence ”Johnny is ill” is not always false, it is equal to the information of ”Johnny has a flu and Johnny is ill” minus the information in the sentence ”Johnny is ill.”

Using the notion of conditional information, we define independence of two sets in the following manner:

1.14. A is independent of $B \Leftrightarrow$ I(A/B) = I(A).

If for example the information in the sentence ”Johnny is ill provided that primroses have blossomed today” is equal to the information in the sentence ”Johnny is ill,” then the sentences ”Johnny is ill” and ”Primroses have blossomed today” are independent.

We supplement the description of the function of information provided in points 1.1., 1.2., 1.3., 1.4., 1.5., 1.6., 1.7., 1.12. and 1.13. by the axiom specifying the information contained in the sum of both sets:

For A, $B \in S$, $A \cap B = \emptyset$
1.15. $\mathrm{I}(A \cup B) = F(\mathrm{I}(A), \mathrm{I}(B))$

where F is a real positive function defined on $R^+ \times R^+$, and therefore F: $R^+ \times R^+ \rightarrow R^+$.

Axiom 1.15. plays a particularly material role by determination of the measure of information. Depending on the choice of function F we obtained varying measures of information. Detailed discussions concerning the properties of function F are presented in the articles of J. Kampé de Fériet, C. Bertoluzza and M. Schneider (1974), C. Langrand (1974) and others. Therefore, I refer all of the readers more interested in this problem to their articles. In this paper I assume that for A, $B \in S$ such that $A \cap B = \emptyset$:

1.16. $F(\mathrm{I}(A), \mathrm{I}(B)) = — c \log \, [e^{-\frac{I(A)}{c}} + e^{-\frac{I(B)}{c}}]$,

where c is a positive constant, which makes it possible to select the unit of information. From 1.15. and 1.16. we get:

1.17. $\mathrm{I}(A \cup B) = — c \log \, [e^{-\frac{I(A)}{c}} + e^{-\frac{I(B)}{c}}]$.

3. MEASURE OF INFORMATION

The secret of success enjoyed by the notion of information lies not only in its generality, thanks to which it may be used by a description of various kinds of events, but above all in fact, that it has been included in the category of measureable values.

The first step in the development of the contemporary theory of information was the quantitative definition of information by means of the measurement of probability. A combination of the notion of information with the notion of probability, and therefore application for the description of information of the notional apparatus of the probability theory, has resulted in the use of the notion of information, when the notion of probability may be used (Shannon, Weaver 1949).

One may also measure information without the measure of probability. I adopt this point of view in this paper.

We shall obtain a unit of information by the introduction of the normalization of function *I*. Let $\Omega = \{A,\ B\}$, $A \cap B = \emptyset$. If $\mathrm{I}(A) = \mathrm{I}(B)$, then in accordance with 1.17. and 1.7.:

2.1. — $c \log 2e^{-\ \frac{I(A)}{c}} = 0$

and therefore

2.2. — c log2 + $\mathrm{I}(A) = 0$ 0

and therefore 0

2.3. $\mathrm{I}(A) = c$ log2 0

If we therefore assume that $c = \ \frac{1}{\log 2}$, then:

2.4. $\mathrm{I}(A) = 10$

The unit of information defined in this manner shall be called a bit. In other words, the information of event A is equal to 1 bit, if and only if set Ω is a two-element set, $\Omega = \{A,\ B\}$, and when $\mathrm{I}(A) = \mathrm{I}(B)$. Generally, when set $\Omega = \{A_1,\ A_2,\ ...,\ A_n\}$ and $\mathrm{I}(A_1) = \mathrm{I}(A_2) = ... = \mathrm{I}(A_n)$, then:

2.5. $J(A_i) = \log_2 n \quad i = 1,\ 2,\ ...,\ n$

For example, when set Ω contains only two elements 0 and 1 such that $0 \cap 1 = \emptyset$ then set $S = \{0 \cap 1,\ 1,\ 0,\ 0 \cup 1\}$ is a check. If we assume that $\mathrm{I}(0) = \mathrm{I}(1)$, then

$\mathrm{I}(0 \cap 1) = +\infty$
$\mathrm{I}(0 \cup 1) = 0$
$\mathrm{I}(1) = 1$
$\mathrm{I}(0) = 1$

Another kind of example is a set of contradictory sentences $\{p,\ {\sim}p\}$. Set $S = \{p \cap {\sim}p,\ p,\ {\sim}p,\ p \cup {\sim}p\}$ is a check. If we assume that $\mathrm{I}(p) = \mathrm{I}({\sim}p)$, then:

$\mathrm{I}(p \cap {\sim}p) = +\infty$
$\mathrm{I}(p) = 1$
$\mathrm{I}({\sim}p) = 1$
$\mathrm{I}(p \cup {\sim}p) = 0$

For example, a measuring device, which reacts only to the colour red, provides one-bit information.

Let us consider the situation, when we have n sentences, each of which may by either false — F or true — V. In other words, we assume that the set of valuations is two-element and that the set of sentences is analysed from the point of view of their value. Then the information of each n obtained from set $\{V, F\}$ is equal to $n \log_2 2 = n$ (Giedymin 1964). In the first multi-value system proposed by Łukasiewicz, the set of valuations is as follows: $\{0, \frac{1}{2}, 1\}$. If we assume that $\mathrm{I}(0) = \mathrm{I}(\frac{1}{2}) = \mathrm{I}(1) = \mathrm{I}(A)$ then $\mathrm{I}(A) = \log_2 3$. If we consider the set of valuations of n sentences, then the information of an n ordered from the three-element set $\{1, \frac{1}{2}, 1\}$ is equal to $\log_2 3^n = n \log_2 3$.

In many mathematical machines, a memory cell is composed of 24 elements, each of which can have the value of 0 or 1. Set Ω is composed of 2^{24} elements. Information of the entire machine ”word” is equal to 24 bits.

In order to write down in the telegraphic code a letter of the Latin alphabet, which has 32 characters, it is sufficient to dispose of five elements of the memory cell, since $\log_2 32 = 5$ bites.

One more example:

In written texts the yes/no question is a one-bit question, In other words, the answer to a yes/no question can provide one bit of information. For example, one answer to the question ”Did John go to school” with either yes or no.

A yes/no question may be understood in a specific context as referring not to the entire expression, but to particular elements thereof. Then, the information of a yes/no question depends on the number of those elements, For example the question ”Did John go to school” may be understood in the following manner (Marciszewski 1974: 133; Koj 1972: 23):

”Did John go to school?” (John? not John?)
”Did John go to school?” (did he go? did he not?)
”Did John go to school?” (to school? not to school?)

The set of elementary events Ω is composed of 8 elements in this case. By the assumption that each answer is equally informative, the answer to the yes/no question provides, in this understanding, 3 bits of information.

Let us return once again to the example with the letters of alphabet. Let us assume that the alphabet of the Polish language has 32 letters and that each of those letter is equally informative. Then $\Omega = \{a, b, ..., x, y, z\}$, $\mathrm{I}(a) = \mathrm{I}(b) = ... = \mathrm{I}(z) = \log_2 32 = 5$ bits.

If we are estimating the information of two letters of the alphabet, then the set of elementary events is a set of ordered pairs of $\Omega \times \Omega$ and contains 32^2 elements. Information of each of the ordered pairs, by assumption of informational equality, is equal to $\log_2 32^2 = 2 \log_2 32 = 2 \cdot 5 = 10$. Respectively, sequences of three, four and five ordered letters of the alphabet may provide information: $3 \log_2 32$, $4 \log_2 32$, ..., $n \log_2 32$.

At this point it is worth noticing that by the same length of a sequence of signs, information contained in such a sequence depends on the set, to which the signs in this sequence belong. For example a sequence of four ordered letters of alphabet $\{0, 1\}$ can at most provide four-bit information ($4 \log_2 2 = 4$), whereas a sequence of four ordered signs being Chinese ideograms, by the assumption that there are 32,768 such ideograms, provides 60 bits of information ($4 \log_2 32{,}768 = 4 \cdot 15 = 60$).

On the basis of formula 1.17. we calculate the information contained in the events, which are the sum of elementary events. A simple example will explain this procedure. Let $\Omega = \{A_1, A_2, B\}$, $\mathrm{I}(A_1) = \mathrm{I}(A_2) = \mathrm{I}(B) = \log_2 3$. Information $(A_1 \cup A_2)$ is calculated in the following manner:

2.6. $\mathrm{I}(A_1 \cup A_2 \cup B) = 0 = — c \log \, [e^{\frac{-I(A1 \cup A2)}{c}} + e^{\frac{-I(B)}{c}}]$

2.7. $e^{\frac{-\ I(A1\ \cup\ A2)}{c}} + e^{-\log 3 = 1}$

2.8. $e^{\frac{-\ I(A1\ \cup\ A2)}{c}} = 1 - e^{-\log\ 3}$

because $e^{-\log\ 3} = \frac{1}{3}$

2.9. $e^{\frac{-\ I(A1\ \cup\ A2)}{c}} = 1 — \frac{1}{3} = \frac{2}{3}$

Logarithmising both sides, we get:

2.10. $— \frac{I(A1\ \cup\ A2)}{c} = \log 2 — \log 3$

2.11. $\mathrm{I}(A_1 \cup A_2) = c \log 3 — c \log 2$

If we assume that 1 bit is our unit, then $c = \frac{1}{\log 2}$ and having put this value in formula 2.11, we get

2.12. $\mathrm{I}(A_1 \cup A_2) = \log_2 3 — \log_2 2$

For example, in the text ”Jaś poszedł do szkoły”(*John went to school*) there are 18 signs plus three spaces, 21 signs in total. The set of elementary events is therefore composed of 21 elements. Information provided by each element of this text is equal to log21. If we want to calculate information of letter *o*, which appears in the text in three different places, then we apply formula 2.12 I($o_1 \cup o_2 \cup o_3$) = $\log_2 21$ — $\log_2 3$. Information contained in letter *d* in the same text is equal to I($d_1 \cup d_2$) =$\log_2 21$ — $\log_2 2$.

Apart from the information of a particular event, we often determine the average information in subsets of set S. Let $\Omega = \{a,\ b,\ c\}$, then $S = \{\{a\},\{b\},\ \{c\},\ \{a,\ b\},\ \{a,\ c\},\ \{b,\ c\},\ \{a,\ b,\ c\},\ \emptyset\}$. The divisions of set S, which we shall mark as P_1, P_2, P_3, P_4, P_5 are as follows:

P_1= $\{\{a\},\{b\},\ \{c\}\}$
$P_2 = \{\{a\},\{b,\ c\}\}$
$P_3 = \{\{a,b\},\ \{c\}\}$
$P_4 = \{\{a,\ c\},\ \{b\}\}$
$P_5 = \{\{a,\ b,\ c\}\}$

The average information for each of these subsets we mark respectively, as I$_{P1}$, I$_{P2}$, I$_{P3}$, I$_{P4}$, I$_{P5}$ and we calculate it in accordance with the following formula:

$$2.13.\ \mathrm{I}_p = \textstyle\sum_{i=1}^{k} I\left(A_i\right) \cdot 2^{-I(A_i)}$$

where k is the number of elements of a given subset.

example:

$$\mathrm{I}_{P1} = \textstyle\sum_{i=1}^{3} I\left(A_i\right) 2^{-I(A_i)} = 3 \cdot\ \ I\left(A_1\right) 2^{-I(A_1)}$$

since I(A_1) = I(A_2) = I(A_3) = $\log_2 3$

$$\mathrm{I}_{P1} = 3{\cdot}2^{-log_2 3}\ \log_2 3 = \log_2 3$$

Division P_1 is the finest division of set S. Information of the finest set is, in the case of average information, equal to the information of the elementary event.

$$\mathrm{I}_{P2} = \mathrm{I}_{P3} = \mathrm{I}_{P4} = (\log_2 3 - \log_2 2) \cdot 2^{\log_2 2 - \log_2 3} + \log_2 3 \cdot 2^{-\log_2 3} = \tfrac{2}{3}\log_2 \tfrac{3}{2} + \tfrac{1}{3}\log_2 \tfrac{3}{1} = -\tfrac{2}{3}\log_2 \tfrac{2}{3} - \tfrac{1}{3}\log_2 \tfrac{1}{3} = \log_2 3 - \tfrac{2}{3}$$

$$\mathrm{I}_{P1} > \mathrm{I}_{P2}$$

The thicker the division of set S, the smaller the average information. It might be said that by a thicker division we are losing information, In case of the thickest division, this is division P_5 in this case, the average information is equal to zero.

4. LANGUAGE AND TEXT

In this paper I use the notion of language, which will be discussed in more detail in the article titled ”O pewnym modelu języka” (Starosta 1974: 147). Let V be a finite not empty set called the vocabulary. The elements of the vocabulary shall be called words, and each finite sequence of words shall be called a phrase. K. Bogacki, the translator of John Lyon’s *Introduction to Theoretical Linguistics,* introduces the term ”rządek” (order) corresponding to the term phrase. In certain publications, translations from Russian, we may also encounter the term ”łańcuszek” (chain), as description of the notion of phrase.

The set of all phrases is the universal language which we will mark as V^*. We say that set V^* is generated by set V. Each subset of set V^* is called a language, of which V is the vocabulary. The language is marked as L.

If we assume that the elements of vocabulary V are three signs $V = \{a, b, c\}$, then the infinite set of phrases which are possible to be created from these three signs, is a universal language generated by V, and each subset, e.g. $\{ab, ac, aaa\}$, $\{bbbb, acacb, a\}$ ... etc. is a language, whose vocabulary is set $\{a, b, c\}$.

If we assume that the elements of the vocabulary are simple expressions of the Polish language, and we will mark as Fa set of these phrases appearing in the epic poem *Pan Tadeusz,* then this set will be a language, whose vocabulary is the set of simple expressions of the Polish language. This is the language of this precise selected epic poem.

Generally, we will describe as language L an ordered pair of $<V, F>$, where V is a non-empty, finite set, and F is a set of selected phrases.

Text shall mean each subset of set F, in other words, each subset of a given language L. Text in this sense means both one-elements of subset F, as well as the entire subset F. In the examples presented above, a text is the set $\{ab\}$, $\{ab, aaa\}$ and the entire set $\{ab, ac, aaa\}$, if we assume that the selected phrases are *ab, ac* and *aaa.* A text is both book four of *Pan Tadeusz*, as well as the sentence ”A już deszcz wciąż pluszczy.”

The same inscription in the Polish language may be studied as a text of the language whose vocabulary is the set of the signs of alphabet, or as a text of the language, whose vocabulary is the set of simple expressions. In the first case, we will be speaking of an analysis at the level of letters, in the second case we will be speaking of an analysis on the level of expressions. Hereinafter, I limit my considerations to these two levels.

It needs to be indicated that in the case of both of these analyses, one needs to clearly define the vocabulary V. E.g. by the letter analysis one needs to decide, whether the considered vocabulary is the Polish alphabet containing 35 signs, or an extended alphabet including all punctuation marks and the most common abbreviations, containing 68 elements, or finally a vocabulary containing only the small letters of the alphabet, without letters *ś, ć, ę, ą, ź, ó,* etc., which contains only 26 signs. We may also assume that vocabulary V contains only the signs which appear in the studied inscription. E.g. if the studied text is the word *kot* (*cat*), then the word *kot* is the only selected phrase of interest for us. The vocabulary V contains only tree signs $V = \{k,\ o,\ t\}$.

By expression analysis it needs to be determined, whether vocabulary V contains almost all simple expressions of the Polish language appearing in Polish literature, or only those appearing in the natural language or in the technical literature, or in a single publication only. By determination of vocabulary V we somewhat determine are measuring apparatus, which we will then apply for text analysis.

The knowledge of this measuring apparatus is of particular importance, if we use the notion of information for the text analysis, since the function of information is the function of the measuring apparatus, in the widest possible understanding of the latter expression. In other words, the measuring apparatus determines the set of information, which is possible to obtain with the use thereof. For example, if we used for the analysis of a written text an apparatus reacting only to red colours, we would obtain only information concerning the colour of the inscriptions, whether they are red or not. If we used for the analysis an apparatus which reacts solely to the length of particular, simple expressions, i.e. the number of letters in an inscription from one space to the next, then we would obtain information concerning the length of the expression in the text. If we limited the text analysis to the abovementioned letter and expression levels, our first task would be to provide an informative description of the vocabulary. Only after we have obtained such a description, is it possible to compare the information contained in the texts in a given, determined language.

Providing a description of the written Polish language requires the use of a mathematical machine, since firstly the use of a machine makes it possible to examine the large empirical material in a relatively short period of time. In Poland machine research has been applied for phonetic analysis. i.a. one has determined the phonetic information contained in the Polish spoken text, in particular the information conveyed by particular phonemes or phonetic dyads, triads or tetrads, on the basis of tape recorded material composed of 100,000 phonemes of running text (Jassem 1974). Information research of written language is carried out sporadically and it is even difficult to speak of a partial description of the information conveyed by the Polish language.

All research carried out in Poland employs the probability theory: one calculates information by measuring the probability of events. The English and Russian languages have been analysed in a similar manner. In this paper I present a proposal of the determination of various types of information contained in text, without resorting to calculation of probabilities. Application of a general theory of information considerably simplifies the calculation procedure and provides direct data concerning the information description of a given language, which we need the most.

In view of the lack of results of measurement of information conveyed by the letters of the Polish alphabet, not mentioning ordered pairs or threes, as well as in view of the total lack of information description of the simple expressions of the Polish language, in the chapter to come I will use very simplified examples, which are sufficient to present the mere research method, but however, do not provide reliable numerical results. The latter could be obtained, as I have mentioned already, after examination of a large portion of empirical material with the use of a mathematical machine.

5. THE NOTION OF INFORMATION AND THE TEXT ANALYSIS

Text analysis from the point of view of information theory requires determination of a set of elementary events, and then a set of events S, for which we will determine the function of information. I am assuming that vocabulary V will be treated by us as a set of elementary events. In case of letter analysis set $\Omega = V$ and contains the signs of the alphabet and other signs comprising vocabulary V. In the case of expression analysis set Ω contains selected simple expressions. By the assumption of equal information of each of the elements of the vocabulary, the average information of the entire vocabulary, as well as the average information of the particular elements thereof, is calculated with the use of formula 2.5.: for each $x \in V$, $\mathrm{I}(x) =$

$\mathrm{I}_a(V) = \log_2 n$, where n is the number of elementary events. For example, if vocabulary V is composed of 64 elements, the average information contained in this vocabulary and the information of each element thereof is equal to $\log_2 64 = 6$ bits. If we assume that the set of elementary events is a vocabulary containing 32,768 simple expressions of the Polish language, then, by assumption of the equal informational value of each of those expressions, the average information of this vocabulary and the information of each of the expressions thereof shall be equal to $\log_2 32{,}768 = 15$ bits.

Average information, equal in the first example to 6 bits, and in the second example to 15 bits, is the maximum average information conveyed by a given vocabulary, which we obtain, if we are dealing with the finest division of vocabulary V. Every thicker division of the vocabulary results in a smaller value of the average information. This is intuitively understandable. We treat vocabulary V, as I have already mentioned, as a measuring device. In case of the finest division of this vocabulary, we get the maximum number of objects, to which a given device reacts. A letter vocabulary divided into individual signs distinguishes every one of these signs as a separate object. The same vocabulary divided for example into vowels and consonants does not distinguish between such signs as *o, i, a* or *e*, neither does it distinguish between signs *g* and *f* or *z* and *c*. Signs included in each of the abovementioned classes shall be undistinguishable. Division into vowels and consonants somewhat combines particular letters with one another, and as a result we obtain a two-element set $P = \{A, B\}$, where A is the set of vowels and B is the set of consonants. Whereby the information contained in set A is not equal to the information contained in set B. If we assume that vocabulary V contains 32 letter of the alphabet, and there are 6 vowels among them, and the remaining letters are consonants, then:

$$\mathrm{I}(A) = \log_2 32 - \log_2 6 = 4 - \log_2 3$$
$$\mathrm{I}(B) = \log_2 32 - \log_2 26 = 4 - \log_2 13$$

The average information of this division $\mathrm{I}(P)$ is therefore equal, according to formula 2.13., to:

$$\mathrm{I}(P) = - 6/32 \log_2 6/32 - 26/32 \log_2 26/32 = 0.7$$

In the above considerations we have assumed the informational equality of the expressions of vocabulary V. Languages used in practice demonstrate an informational variety of the vocabulary elements. In order to obtain the value of the information conveyed by particular expressions of the dictionary, one needs to examine a certain specific number of texts of the Polish language, sufficiently large enough to be treated as a representative sample. For example, if the sentence ”Jaś poszedł do szkoły” was our representative text, then the information conveyed by the letter *o* in this text is equal to $\log_2 18$ — $\log_2 3$, information conveyed by letter *a* is equal to $\log_2 18$, information conveyed by letter *ł* is equal to $\log_2 18$ — $\log_2 2$, etc. In the text ”Dziś zakwitły krokusy” (Crocuses have bloomed today) information conveyed by letter *o* is equal to $\log_2 19$ — $\log_2 2$, and the information conveyed by the letter *ł* is equal to $\log_2 19$.

In a similar manner we can calculate the information contained in pairs, threes, fours or fives of ordered letters of a given alphabet. The number of signs in the examination sample increases thereby.

It is much more troublesome to calculate information in case of text analysis on the expression level. If we assume that the set of elementary events has 30,000 elements, which is not a lot, then if we are to consider the simple expressions of the natural language, then the set of pairs of theses expressions shall have $30{,}000^2$ elements and the set of threes shall have $30{,}000^3$ elements. Not all twos and threes appear in the natural language. However, in order to determine the information contained in the pairs or threes, one needs to examine texts counting hundreds of thousands of expressions. This task can be accomplished only with the use of a mathematical machine.

Informational text analysis may be performed partially in selected languages, e.g. we may consider the texts of the language of chemistry or fragments of publications on farming equipment, etc. The considered vocabularies contain in such cases only the expressions which can be found in such publications. A vocabulary obtained on the basis of such an initial analysis, in particular a frequency vocabulary, is used as a template, with which we compare the vocabularies obtained from the analysis of other texts from the same field and the vocabularies obtained from the analysis of texts from other fields. Comparing vocabularies makes it possible, for example, to find the so-called specific terms of a given field or particular publications. Yet, I do not intend to discuss this problem in any detail.

At the end I would like to note that it is possible to use the information text analysis in other kinds of ways, which are particularly interesting from the point of view of semiotics. It seems namely that by using extremely simple assumptions concerning the notions of information, language and text, which I have presented in the preceding chapters, it is possible to obtain a relatively rich description of syntactic, semantic and pragmatic relations of the natural Polish language. Justification of this claim requires a separate paper, in which one needs to i.a. provide results of particular machine examinations. In this work I merely signal the problem by presenting a simple example, which gives a certain idea of both the research method, as well as the issue itself.

The first text that we are going to analyse is as follows: ”Wartość istnieje niezależnie od człowieka. Wartość istniała przed człowiekiem” (”Value exists irrespectively of the human. Value existed before the human did”). Vocabulary V_1contains the following elements: $V_1 = \{$Wartość, istniała, istnieje, niezależnie, od, człowieka, człowiekiem, przed$\}$. For the sake of simplification we assume that the piece of information carried by each of the words from the vocabulary is equal with respect to quantity: the piece of information carried by the words of the vocabulary is equal to the average information and is equal to $\log_2 8 = 3$ bits.

Let us consider in turn a set of ordered pairs, threes, fours, fives, sixes, sevens and eighths of vocabulary V_1. In Polish only certain combinations of expressions are treated by the users of this language to constitute correctly constructed and meaningful sentences. And so, for example, from among 64 possible ordered pairs, the following pairs can be considered to be sentences in the Polish language: <wartość istnieje>, <istnieje wartość>, <wartość istniała>, <istniała wartość>, <istnieje niezależnie>, <niezależnie istnieje>, <istniała niezależnie>, <niezależnie istniała>. Such pairs, as for example, <istnieje człowieka>, <wartość od> or <wartość człowieka> are not considered sentences in the Polish language and therefore we shall reject such pairs. The selection is obviously arbitrary to a certain extent. In each case, not all pairs are equally informative. The piece of information contained in the pairs calculated on the basis of an experiment is smaller than the piece of information calculated theoretically, which is equal to $2 \log_2 8 = 6$ bits. The value of information in the case of the text presented above calculated on the basis of an experiment is equal to $\log_2 8 = 3$ bits.

Theoretically, the informational value in the presented example is equal to $3 \log_2 8 = 9$ bits. In practice, we consider the following threes to be meaningful sentences of the Polish language: <istnieje wartość niezależnie>, <niezależnie istnieje wartość>, <wartość istniała niezależnie>, <niezależnie istniała wartość>, <istniała przed człowiekiem>, <przed człowiekiem istniała>, <istnieje przed człowiekiem>, <przed człowiekiem istnieje>. The informative value of the threes is equal to $\log_2 10$.

Information contained in the fours is equal to $4 \log_2 8 = 12$ bits. We consider the following sentences to be meaningful: <wartość istniała przed

człowiekiem>, <istniała wartość przed człowiekiem>, <przed człowiekiem istniała wartość>. We moreover consider meaningful the three sentences in which <istniała> is replaced with <istnieje>, as well as the 6 sentences obtained from combining various expressions in the sentences: <istnieje niezależnie od człowieka>, <istniała niezależnie od człowieka>. In total, we assume that it is possible to obtain 12 meaningful four-word sentences from vocabulary V_1. Information contained in the fours is equal to $\log_2 12$.

Information contained in the fives is theoretically equal to $5 \log_2 8 =$ 15 bits. In practice, we consider the following sentences to be meaningful: <Wartość istniejenie zależnie od człowieka>, <wartość istniała niezależnie od człowieka>, <wartość istniała niezależnie przed człowiekiem>, and sentences obtained from the possible mix of expressions in the above sentences, which results in a total of 12 sentences. Information contained in the fives is equal to$\log_2 12$.

We do not in practice treat ordered sixes and eights obtained from this dictionary as sentences, but we do consider sevens as sentences. These are the sentences composed of combinations of expressions <wartość istniała przed człowiekiem niezależnie od człowieka> and <wartość istnieje przed człowiekiem niezależnie od człowieka>, which in total gives us 16 sentences. The information contained in the sevens is equal to $\log_2 16 = 4$ bits. Theoretically, the information of the sevens is equal to $7 \log_2 8 = 21$ bits.

We will compare with the above text another text which is also composed of eight elements: ”Błysneło. Kobieta stanęła. Tygrys podchodzi powoli. Dziecko zapłakało” (It flashed. A woman stopped. A tiger is approaching slowly. A baby cried.). Vocabulary V_2 contains the following elements: $V_2 =$ {błysnęło, kobieta, stanęła, tygrys, podchodził, powoli, dziecko, zapłakało}. The average information value of the vocabulary is 3 bits. Theoretically, information contained in ordered pairs, threes, fours, fives and sixes is equal, as in the previous example, to 6, 9, 12, 15 and 18 bits respectively. In these examples, however, different numbers specify the information contained in the pairs, threes and fours calculated on the basis of an experiment. Information contained in pairs, if we consider the following pairs to be meaningful sentences <błysnęło powoli> and <dziecko błysnęło>, is equal to $\log_2 14$, and if we reject these pairs, is equal to $\log_2 10$. Information contained in threes is equal at most to $\log_2 9$. Fours and fives cannot be obtained from the given vocabulary.

Juxtaposition of both texts provides the following results:

Text I	Text 2
I(V_1) $= \log_2 8 = 3$ bits	J (V_2) $= \log_2 8 = 3$ bits
I(pairs) $\log_2 8$	$\log_2 10$
I(threes) $\log_2 10$	$\log_2 9$
I(fours) $\log_2 12$	—
I(fives) $\log_2 12$	—
I(sixes) —	—
I(sevens) $\log_2 16$	—

The first conclusion that comes to mind concerns the elements of both vocabularies: the elements of the first vocabulary less freely combine in pairs than the elements of the second vocabulary. Comparing to vocabulary V_2 however, their combinations in threes, fours, fives and even sevens are less limited. The elements of vocabulary V_2 are in a sense predetermined as far as the possible combinations are concerned. As a result, text 2 is far less informative than text 1. The question now is, where the differences come from.

A comprehensive answer to this question cannot be obtained without conducting competent research. We may suspect however that we will not get such an answer if we do not enrich the language model. As a result of the research we obtain, however, a material which only partially specifies the differences between the theoretically calculated and practically used information characteristic for a given text. These differences indicate the existence of certain regularities. We may assume that the difference as to the amount of information theoretically carried by a text and the amount of information possibly carried is indirectly the measure of these regularities — it is the measure of the semiotic information of the text. One should expect that the greater the differences, the more information contained in syntactic, semantic and pragmatic relations of a given text.

It is difficult for me to say at the moment, whether, if we stayed at the research level discussed above, it would be possible to experimentally separate particular types of semiotic relations. Separating syntactic relations from semantic relations seems doable: we namely calculate experimentally the set information of pairs, threes, fours, etc., obtained from a given vocabulary, being directed firstly by the choice of sentences by their mainly syntactic correctness (by provision of strict grammatical rules, etc.) and secondly by their meaningfulness (meaning). The obtained difference in the layer of information would characterise the semantic information. Should it turn out that the difference is practically equal to zero, we could say that the regularities occurring in the language are of syntactic character.

If we wanted to specify the syntactic and semantic information of a text in more detail, and later also separate and calculate the pragmatic information, one would need to enrich the theoretical language model. One of the steps in this direction is the following:

Let us assume that the language is an ordered three $<V,\ F,\ P>$, where P is a set of divisions of set V (Starosta 1974). Among the possible divisions we differentiate e.g. the division into distribution classes and the division into paradigms. In case of the division of vocabulary V_1 into distribution classes, we get the following set:

P_{D1} = {{wartość}, {istnieje}, {istniała}, {niezależnie}, {od}, {człowieka}, {przed}, {człowiekiem}}

In case of division of vocabulary V_1 into paradigms, we obtain the following set:

P_{P1} = {{wartość}, { istniała, istnieje}, {niezależnie}, {od}, {przed}, {człowieka, człowiekiem}}

In case of division of vocabulary V_2 divisions into distribution classes and paradigms are as follows:

$P_{D2} = P_{P2}$ = {{błysnęło}, {kobieta}, {stanęła}, {tygrys}, {podchodził}, {powoli}, {dziecko}, {zapłakało}}

Vocabulary V_1 constitutes from this point of view a less precise measuring device than V_2. Information contained in divisionP_{P1} is smaller than the information contained in division P_{P2} and smaller than the information contained in vocabulary V_1. In this vocabulary certain expressions may be combined with one another in accordance with the division rules. Information contained in the division rules limits the possibilities of the free combining of expressions of the vocabulary: reduces the information contained in the vocabulary itself.

Summing up: text information analysis may be carried out, if we determine the langue to which we include a given text and if we provide the information description of this language, the Information description of the language is a blueprint with which we compare information descriptions of particular texts.

Information provided by the language is specified on various levels.

In particular this is the level of letters and expressions. Examination of the information provided by a given language on the level of expressions results in the setting of curtain regularities, which result in reduction of the theoretically possible information contained in the texts. Information contained in the text should be treated as locating information: it decreases to zero, when the searched object is placed in the area of events. This locating information is the semiotic information.

It seems that determination of the information of text will make it possible to describe the syntactic, semantic and pragmatic relations of the language in the language of information theory.

Bibliography

1. Barr-Hillel, Yehoshua and Rudolf Carnap (1953) ”An Outline of the Theory of Semantic Information.” *British Journal for the Philosophy of Science* 4: 147-157.

2. Domotor, Zoltan (1970) ”Qualitative Information and Entropy Structures.” In *Information and Inference*, Jaakko Hintikka and Patrick Suppes (eds.), 148-194. Dordrecht: Reidel.

3. Forte, Bruno and Nicolo Pintacuda (1968) ”Information fournie par une expérience.” *Comptes Rendus de l’Académie des Sciences, ser. A* 266: 242-245.

4. Giedymin, Jerzy (1964) *Problemy, założenia, rozstrzygnięcia.* Poznań: PWN.

5. Hintikka, Jaakko (1970) ”Surface Information and Depth Information.” In *Information and Inference*, Jaakko Hintikka and Patrick Suppes (eds.). Dordrecht: Reidel.

6. Ingarden, Roman (1963) ”A Simplified Axiomatic Definition of Information.” *Bulletin of the Polish Academy of Sciences, ser. Math, Astr, Phys* 11: 209-211.

7. Ingarden, Roman and Kazimierz Urbanik (1962) ”Information without Probability.” *Colloquium Mathematicum* 9: 131-150.

8. Jassem, Wiktor (1974) *Mowa a nauka o łączności.* Warszawa: PWN.

9. Kampé de Férier, Joseph (1974) ”La théorie généralisée de l’information et la mesure subjective de l’information.” *Lecture Notes in Mathematics* 398: 1-36.

10. Kampé de Fériet, Joseph, Bertoluzza, Carlo and Michel Schneider (1974) ”Information totalement composable.” *Lecture Notes in Mathematics* 398: 90-99.

11. Koj, Leon(1971) ”Analiza pytań II. Rozważania nad strukturą pytań.” *Studia Semiotyczne* 3: 23-39.

12. Langrand, Claude (1974) ”Précapacités fortes et mesure d’information.” *Lecture Notes in Mathematics* 398: 36-49.

13. Losfeld, Joseph (1972) ”Information moyenne dans une épreuve statistique.” *Comptes Rendus de l’Académie des Sciences, ser. A* 275: 509-512.

14. Losfeld, Joseph (1974) ”Information Généralisée et Relation d’Ordre.” *Lecture Notes in Mathematics* 398: 49-62.

15. Marciszewski, Witold (1974) ”Analiza semantyczna pytań jako podstawa reguł heurystycznych.” *Studia Semiotyczne* 5: 133-146.

16. Marcus, Solomon (1970) *Teoretiko-mnozhestviennye modeli yazykov.* Moskva.

17. Pintacuda, Nicolo (1969) ”Prolongement des measures d’information.” *Comptes Rendus de l’Académie des Sciences, ser. A* 269: 861-864.

18. Sallantin, Jean (1974) ”Information et Trajectories sure un systéme de Propositions.” *Lecture Notes in Mathematics* 398.

19. Shannon Claude E. and Warren Weaver (1949) *The Mathematical Theory of Communication.* Urbana: University of Illinois Press.

20. Starosta, Barbara (1974) ”O pewnym modelu języka naturalnego.” *Studia Semiotyczne* 5: 147-157.

21. Urbanik, Kazimierz (1972) ”On the Concept of Information.” *Bulletin of the Polish Academy of Sciences,ser. Math, Astr, Phys.* 20: 887-890.

22. Varma, Suneeta and Prem Nath (1967) ”Information Theory — A Survey.” *Journal of Mathematical Sciences* 2: 75-169.