Barbara Starosta

CONTRIBUTION TO THE SEMIOTICS OF QUESTIONS

Originally published as "Przyczynek do semiotyki pytań," Studia Semiotyczne 6 (1975), 95–103. Translated by Magdalena Tomaszewska.

Introduction

The issue of the logic of questions was relatively popular with Polish authors (Ajdukiewicz 1960; Giedymin 1964; Kubiński 1971). Works by Tadeusz Kubiński, Kazimierz Ajdukiewicz, Jerzy Giedymin are mentioned by almost all researchers of this issue who publish in English, German or Russian (Åqvist, Harrah, Voishvillo). However, taking into consideration publications about the analysis of artificial and natural languages, the number of works devoted to the semiotics of questions is negligible. On the basis of the bibliography it could be deduced that the issue is either closed and nothing new can be added, or it is an alleged problem and thus not worth spending time on. In the first case, a theory of questions should be developed and verified, while in the other case, there would be no need to create such a theory.

The solution to this dilemma can be sought in the practice of information technology research and the fact that mathematical machines are becoming more popular in scientific research. It turns out that a theory of questions is necessary, for example, to construct theories of information finding systems which are generally based on asking QUESTIONS by a user of a given finding system. Moreover, a theory of questions would be extremely useful in the research on how data are organized in the memory of a machine.

In light of these simple needs, the present state of research on the logical theory of questions is unsatisfactory. The problem is not solved. Hence, Leon Koj (Koj 1971, 1972) focuses on the logical anlaysis of questions in two

articles, and Witold Marciszewski (Marciszewski 1974) also addresses the issue of questions in an article.

The present article is another attempt to face the emerging needs. I shall use a series of assumptions formulated by the above mentioned authors to the full extent. I shall also refer to the "pioneers" mentioned at the beginning.

In relation to the whole of the issue, the scope of my research is extremely narrow. Firstly, I focus on the syntax of questions, and to be more specific — on such a theory of questions in which there are no semantic or pragmatic terms. In my opinion, such a theory plays an analogous role on a negative vocabulary: it clearly separates what can be said about questions exclusively on the basis of the internal structure of expressions, from what can be described only in a language enriched with semantic and pragmatic terms. Secondly, I focus my research only on written language.

1. The notion of question

Before I define the notion of question that I shall use in this article, I shall present characteristics of the language in which I shall distinguish the notion.

Let V stand for a finite set of simple expressions used in Polish writings. It is a set of all NON-ISOMORPHIC inscriptions separated by a space, $V = (a_1, ..., a_n)$. Now, I determine the operation of joining, called concatenation, of set V^* generated by set V. The elements of this set, which I call phrases, are: expressions of vocabulary V, pairs of these expressions joined as a result of concatenation, and all possible complex expressions obtained through repetition of the operation of joining of the mentioned elements.

Among all possible phrases, only part occurs in written Polish texts. These phrases are treated here as DISTINCTIVE PHRASES and their set is marked with ϕ. To such distinctive phrases belong the following inscriptions e.g.: *trawa rośnie wysoko* "the grass grows high," *szczyt głupoty* "the peak of stupidity," *Jan z Czarnolasu* "John of Blackwood," *Bibliografia* "bibliography," etc.

What is referred to as language L is an ordered pair $L = <V, \phi>$. A detailed description of this model of language was presented in my article *O pewnym modelu języka naturalnego* (Starosta 1974).

What is referred to as QUESTION is the phrase in language L which ends in a question mark symbolized with *?*.

Now, I determine the notion of SPECIFIC TERM of question, or in shorter words — INTERROGATIVE TERM, as follows:

Expression a, $a \in V$ is a specific term of question, that is an INTERROGATIVE TERM if and only if it occurs ONLY in questions.

By introducing the above definition of interrogative term I go beyond the framework outlined in language L: for I make a DIVISION of vocabulary V into expressions that occur in questions and expressions that do not. Language L characterized above does not presuppose the operation of division. It is convenient then to extend the notion of language for further considerations.

If P stands for a finite class of divisions of vocabulary V, and $P = (P_1, P_2 \ldots P_n)$, then language L is defined as an ordered set $<V, \phi, P>$.

A division can be made of the vocabulary into expressions that occur only in questions and the remaining ones which differentiate from the vocabulary such as interrogative terms: *czy* "auxiliary DO, BE, HAVE, ...," *który* "which/what," *jaka* "which/ what," *gdzie* "where," *dlaczego* "why," and many others. This is one of the possible divisions of vocabulary, which can be conventionally called division D_1. Also, vocabulary V can be divided into expressions that occur in EVERY distinctive phrase, and expressions that occur only in phrases that end in: a dot, a question mark, or an exclamation mark. This division, which I mark as D_2, can be conventionally treated as a division of expressions of the vocabulary into NAMES and non-names. I shall not introduce further divisions of vocabulary V. Hence, as a result, language L is characterized by vocabulary V, distinctive phrases ϕ, and two divisions D_1 and D_2.

Distinctive phrases which consist of not only names but also expressions that are not names shall be called SENTENCES. In the case of such a division of distinctive phrases, questions belong to the set of sentences. A separate subset in the set of sentences are sentences that end in a dot. I shall call such sentences affirmative.

2. Classification of questions

The most detailed division of questions into types is the division based on the shape of the interrogative term. In the case of such a division, there are as many types of questions as there are interrogative terms in vocabulary V. Other criteria of division are extra-syntactic. For example, when the notion of a set of potential answers is introduced with a reference to conventions of using questions and extra-linguistic knowledge, so that questions are characterized depending on the characteristics of the set of answers.

If we treat interrogative terms as operators binding variables which take values from the sets of answers, then the type of set of answers may be a

basis for question classification. In this case we assume that two types of questions are equal if the sets of their potential answers are equal. This latter type of division is adopted by e.g. Koj who divides questions into two classes: questions in which the interrogative operator binds a name variable, and questions in which the interrogative operator binds a sentence variable. In the case of another criterion of division, when finiteness or infiniteness of sets of answers is taken into consideration, a division into open and closed questions is obtained. There may be many divisions of sets of answers, and hence there may be many possible question classifications.

In this work I shall use a statistic criterion. I divide interrogative terms dichotomically into such that appear most frequently in questions of language L, and the remaining ones. The statistical sieve separates the interrogative term *czy* "auxiliary DO, BE, HAVE, ..." In further considerations I shall only analyze questions in which there is the interrogative term *czy*. I shall call such questions *czy*-questions.

Czy-questions are the most frequent questions in scientific publications. They occur extremely often in texts on the methodology of teaching. They are the main type of question in the so called curriculum-based teaching and all kinds of tests. Their commonness is due to what, by analogy to name-properties, may be called DEFINITENESS. For *czy*-questions are the only ones that determine a definite set of potential answers in a mechanical manner, so to speak, by means of only the notion of sentence negation. Moreover, they entropy of the *czy*-question may be calculated on the basis of analysis of the structure of the question itself. Additionally, *czy*-questions determine not only the quantity but also the quality of the information provided by their answers. They have an analogous role as gauging instruments used in scientific research and practice.

3. Types of *czy*-questions

In many works, questions with the operator *czy* "auxiliary DO, BE, HAVE,..." are called closed end questions or fixed-alternative questions. For Tadeusz Kubiński such questions are only the ones in which the interrogative term *czy* occurs only once and when the set of answers to the question consists of two elements. The set consists of the sentence following the interrogative term and the negation of the sentence. For example, the set of potential answers to the question: *Czy Jaś lubi lody?* ["Does John like ice-cream?"] consists of two sentences: *Jaś lubi lody* ["John likes ice-cream"] and *Nieprawda, że Jaś lubi lody* ["It's not true that John likes ice-cream."] Already at this moment it is worth highlighting that the set of answers is

unequivocally determined by the question. Elements of the set are disjunctive and mutually exhaust all possible answers.

Following Kubiński, I assume that closed end questions are a special case of n-element *czy-* questions when $n = 1$. Kubiński differentiates two types of questions among n-element *czy-* questions: conjunctive and alternative.

Conjunctive questions occur in texts in two equivalent forms. If p_1, p_2, ..., p_n are affirmative sentences, then the notation of the two forms of questions is:

$$Czy\ (p_1 \wedge p_2 \wedge ... \wedge p_n)?\ \text{and}\ Czy\ p_1 \wedge czy\ p_2 \wedge ... \wedge czy\ p_n?$$

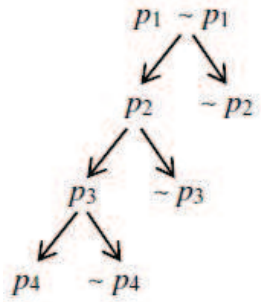
In the case of both forms of conjunctive questions, the set of answers A is: $A = \{p_1 \wedge p_2 \wedge ... \wedge p_n, \sim p_1 \wedge p_2 \wedge ... \wedge p_n, \sim p_1 \wedge \sim p_2 \wedge ... \wedge p_n,$ $..... \sim p_1 \wedge \sim p_2 \wedge \sim ... \sim \wedge\ p_n\}$.

The set of potential answers to a conjunctive *czy*-question consists of 2^n elements, where n is the number of interrogative terms *czy* which may occur in the question.

Alternative *czy*-questions may be presented as follows: $Czy\ (p_1 \cup p_2 \cup ... \cup p_n)$, where $\cup$ stands for a disjunctive alternative. For example, *Czy wyjedziesz w góry, czy nad morze, czy też pozostaniesz w domu*? ["Are you going to go to the mountains, the seaside, or maybe stay at home?"].

The sets of answers to alternative questions are precisely defined. If n is the number of interrogative terms in the question, then the set of potential answers consists of n elements.

Except for conjunctive and alternative *czy*-questions, Kubiński differentiates CONDITIONAL *czy*-questions. They are a sequence of *czy*-questions in which the following question depends on the answers to the previous questions. Questions of this type are presented by means of a tree diagram in many publications. In the simplest case — with one-element *czy*-questions, the conditional *czy*-question has the following form: *Czy jeżeli* p_1 *to* p_2, *a jeżeli* p_2, *to czy* p_3, *a jeżeli* p_3, *to czy* p_4, *itd.* ["If p_1 then p_2, and if p_2, then p_3, and if p_3, then p_4, etc."]. The tree diagram in such a case is as follows:



The *czy*-question tree is frequently used in scientific research. Questions of this type do not determine a straightforward set of answers, but indicate the direction of narrowing the set: they force us to look in a certain set, then in a subset of the set, then in a subset of the subset, etc. until reaching a set of answers that is easier to search through and significantly less numerous than the initial set.

The first *czy*-question and the following ones may consist of n-elements. When *czy*-questions are alternative we deal with an alternative n-element tree, when *czy*-questions are conjunctive, a tree is called conjunctive or multiplicative (Watanabe 1969). I shall not analyze conditional *czy*-questions as they require a separate study.

To sum up, *czy*-questions divide into alternative and conjunctive questions, moreover there is a distinctive group of conditional *czy*-questions. In the case of all these types, the set of potential answers is unequivocally determined by the structure of the question. In order to establish the set it is not necessary to have access to additional knowledge. All the information comes from the question itself.

From the point of view of the user, *czy*-questions are extremely informative, for they give much information about the SET of answers. Simultaneously, they are questions of relatively low information value: in the case of a closed end question, the information maximally amounts to 1 bit. A simple example should explain this apparent paradox, and at the same time will be a starting point to draw an analogy between the role of *czy*-questions in language L and the role of gauging instruments.

If the question is, e.g.: *Jakie jest napięcie w akumulatorze samochodowym?* ["What is the car battery voltage?"], the set of potential answers is determined in very general terms. For it consists of all affirmative sentences which describe all possible car battery voltages. The information contained in the question equals the question's entropy and amounts to hundreds of bits. It is worth reminding ourselves here that if $E_i$stands for the question's initial

entropy, and E_f stands for the question's entropy after receiving the answer, and if we assume that the answer provides exhaustive information, then the question's information equals the remainder of entropies E_i — $E_f = I$, and $E_f = 0$. The question's information equals the question's entropy: $I = E_i$.

If the question *jakie* "what" is supplemented or replaced by the question *czy* "auxiliary DO, BE, HAVE, ...," then the set of answers becomes more defined, e.g. *Czy napięcie akumulatora samochodowego wynosi 2 V, czy 3 V, czy 12 V, czy też 24 V?* ["Is the car battery of voltage 2 V, or 3 V, or 12 V, or perhaps 24 V?"], or *Czy napięcie akumulatora samochodowego wynosi 0, 1/2, 1, ..., 24 V?* ["Is the car battery of voltage 0, 1/2, 1, ..., 24 V?], etc. Admittedly, thus formulated questions still do not inform us which car battery is meant, but the information of these particular questions may be calculated. It is $\log 4 = 2$ bits and $\log 48 < 5$ bits.

When it is said that question A is extremely informative, what is meant is either the difference in the entropy of the question or a question of a different type, e.g. the difference in the entropy of the question *jakie* "what" and the question *czy* "auxiliary DO, BE, HAVE, ..." in the above examples, or what is meant is a situation in which it is assumed on the basis of additional knowledge that one of the expected answers is hardly probable. When, as a result of an experiment or research, this very answer proves to be correct, it is said that it provided much information. For example, there are two potential answers to the question *Czy istnieje życie na Marsie?* ["Is there life on Mars?"], that is: *there is life on Mars* and *there is no life on Mars*. When the element set of potential answers to the question is considered, then the information of the question is 1 bit. However, it is possible to treat each of the potential answers separately and consider the information of each of them, then the information of the answer *there is life on Mars* may be enormous. It depends on our calculation of the probability of this answer. For example, if it is assumed that the probability of this answer amounts to only $\frac{1}{10000}$, then $I =$ — $\log \frac{1}{10000}$ $I = \log 10000 \simeq$ 14 bits.

I shall not discuss the above-mentioned types of information in this article. The first requires an analysis of questions of a different kind than *czy*-questions, the other is related to the pragmatic nature of research. I shall, however, discuss the information of *czy*-questions that is determined inclusively on the basis of the structure of these questions. But before doing so, I shall highlight the analogy between the role of *czy*-questions in language L and the role of gauging instruments.

Let us return to the example of the car battery voltage. In order to learn what the car battery voltage is, it needs to be measured by means

of a proper range voltmeter. The instrument we use determines the set of potential results of measurement. Hence, e.g. a voltmeter with a range from 600 to 1000 V, and a precision voltage reference of 10 V determines a different set of potential answers than e.g. a voltmeter with a range from 0 to 100 V, and a precision voltage reference of 1 V.

In language L, *czy*-questions function as such gauging instruments: they narrow and determine in advance the set of potential answers. The analogy is especially useful when one wants to establish the information of *czy*-questions by means of the notion of information defined without referring to the notion of probability (Ingarden 1963). Then, information is a function of a gauging instrument, or — more precisely — as a function of the set of potential results of measurement. In the case considered here, the information of the *czy*-question is a function of the set of potential answers to the question.

4. Information enclosed in the *czy*-question

I shall not deal with the notion of information here. Those interested should refer to the article "Uwagi o pojęciu informacji" [Some remarks on the notion of information] (Starosta 1973) and works listed in the bibliography. May I recall, however, that information is a function defined on the subsets of non-empty and finite set X. The class of subsets of set X, which is marked here by U, is a Borel field. It is closed under the set-theoretic operations of union and intersection, and, moreover, contains the entire set X treated as one or a certain event, and the empty set treated as zero or an impossible event. The function of information which is defined for the entire complex of events is characterized by the following axioms (Ingarden 1963):

1. If set B is a subset of set A, $(A,\ B \in\ U)$,
   then
   $$I(B) \geqslant\ I\ (A)$$

2. Two sets A and B, $(A,\ B \in\ U)$ are independent if and only if
   $$I(A \cap B) = I(A) + I(B)$$

3. The information of a certain event equals zero

$$I\ (X) = 0$$

4. The information of an impossible event equals $+\infty$

$$I(0) = +\infty$$

Additionally, if we introduce normalization, then:

5. $I(A) = 1$ if and only if

   a. Set A contains two elements, $A = \{a_1,\ a_2\}$
      and
   b. $I(a_1) = I(a_2)$

Introducing the notion of information which is defined by the above axioms enables us to formulate a few theorems concerning *czy*-questions.

THEOREM 1. If a set of potential answers to question B is a subset of potential answers to question A, then the information contained in question B is not greater than the information contained in question A. For example, the information in the question: *Czy kolor tej ściany jest czerwony, czy zielony, czy żółty?* ["Is the colour of this wall red, green, or yellow?"] is smaller than the information in the question *Czy kolor tej ściany jest czerwony, czy zielony?* ["Is the colour of this wall red, or green?"].

A conclusion arises that, in order for the information contained in the two questions A and B to be equal, the condition that the sets of potential answers are equally numerous is not sufficient. These sets must contain THE SAME elements. For example, the information in the question *Czy Zbyszek bawi się, czy poszedł do szkoły?* ["Is Zbyszek playing, or is he at school?"] is not equal, according to this theorem, to the information contained in the question *Czy Jasio leży w łóżku, czy biega po pokoju?* ["Is John lying in bed, or is he running around the room?"]. For, sets of potential answers to each of these questions are disjoint.

However, if we take into consideration the normalization axiom, then the information of the first question is equal to the information of the second question and amounts to 2 bits. In this case, we are only interested in how many elements the sets of answers contain.

THEOREM 2. The information contained in question A and question B equals the sum of information of each of these questions if and only if the sets of potential answers to these questions are independent.

When these sets are dependent, the information contained in both of these questions equals the sum of information of each of these questions diminished by the information provided by the set of answers which is the intersection of both sets considered. For example, the information contained in the questions *Czy Stasio bawi się, czy też poszedł do szkoły?* ["Is Stasio playing, or is he also at school?"] and *Czy Stasio bawi się, czy też leży w łóżku?* ["Is Stasio playing, or is he also lying in bed?"] is treated as equal to the information of the question *Czy Stasio bawi się, czy poszedł do szkoły, czy też leży w łóżku?* ["Is Stasio playing, or is he at school, or is he also lying in bed?"].

The measure of the probability of the set of potential answers to *czy*-questions is determined by definition of information, in the following way (Ingarden, Urbanik 1961):

THEOREM 3: If $I(A)$ is information in the set of *czy*-questions, then for each $A \in U$ there is one and only one positive measure of probability $P(A)$ determined on A and such that for each subset of set A, $B \subset A$ and for each $b \subset B$.

$$P(b/B) = \frac{P(b/A)}{P(B/A)}$$

and

$$I(A) = -\sum_{i=1}^{n} Pa/A \; log \; a_i/A,$$

where a_1, a_2, ..., a_n, stand for all disjoint subsets of A which exhaust set A.

In the case when set A is interpreted as a set of potential answers to a *czy*-question, a_1, a_2, ..., a_n are elements of this set.

Calculating the information of a *czy*-question, we can use the normalization axiom directly, or theorem 3. In the latter case, which uses the assumption that every potential answer to a *czy*-question is equally probable, that is when $P(a_1/A) = P(a_2/A) = ... = P(a_n/A)$, $P(a/A)$ equals $\frac{1}{n}$, then:

$$I(A) = \log n,$$

where n is the number of sentences in the set of potential answers.

Finally, it is worth remarking that the information of n-element alternative *czy*-questions is in principle smaller than the information of n-element conjunctive *czy*-questions. For the set of answers to alternative questions amounts to n elements, while the set of answers to conjunctive questions amounts to 2^n elements. This explains the intuition connected with such questions: conjunctive *czy*-questions are less determined and, with the same assumptions, allow more possible answers than alternative questions.

To sum up: I have shown that the information of *czy*-questions is determined by their structure. Moreover, I have drawn the conclusion that in order for two pieces of information to be EQUAL, the condition that the sets of potential answers are equally numerous is not sufficient, the sets must be identical. Two sets can provide the same quantity of information which is qualitatively different. Last but not least, I have remarked that alternative *czy*-questions are less informative than conjunctive *czy*-questions when they have the same number of elements.

Bibliography

1. Ajdukiewicz, Kaziemierz (1960) "Zdania pytajne." In *Język i poznanie*, vol. 1, 278-286. Warszawa: PWN.

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

3. Ingarden, Roman (1963) "A Simplified Axiomatic Definition of Information." *Bulletin of the Polish Academy of Sciences (Mathematics, Astronomy and Physics Series)* 11: 209-212.

4. Ingarden, Roman and Kazimierz Urbanik (1961) "Information as a Fundamental Notion of Statistical Physics." *Bulletin of the Polish Academy of Sciences (Mathematics, Astronomy and Physics Series)* 9: 315-316.

5. Koj, Leon (1971) "Analiza pytań. Problem terminów pierwotnych logiki pytań." *Studia semiotyczne* 2: 99-113.

6. Koj, Leon (1972) "Analiza pytań. Rozważania nad strukturą pytań." *Studia semiotyczne* 3: 23-39.

7. Kubiński, Tadeusz (1971) *Wstęp do logicznej teorii pytań*. Warszawa: PWN.

8. Marciszewski, Witold (1974) "Analiza semantyczna pytań jako podstawa reguł heurystycznych." *Studia semiotyczne* 5: 133-146.

9. Starosta, Barbara (1973) "Uwagi o pojęciu informacji." *Studia semiotyczne* 4: 95-107.

10. Starosta, Barbara (1974) "O pewnym modelu języka naturalnego." *Studia semiotyczne* 5: 147-157.

11. Watanabe, Satoshi (1969) *Knowing and Guessing. A Quantitative Study of Inference and Information*. New York: Wiley.