Olgierd Adrian Wojtasiewicz

AN ATTEMPT AT FORMALISING A DEFINITION OF THE NOTION OF A 'SUMMARY'[1]

Originally published as "Próba formalnej definicji pojęcia streszczenia," Studia Semiotyczne 7 (1977), 185–191. Translated by Julita Mastelarz.

The notion of a summary is universally known and used by documentalists, yet it does not seem to have a precise definition, let alone one that would comply with formal requirements. Which is not to mean that, due to its intuitive character, the interpretation of the term 'summary' or its use causes major confusion. To the contrary — the relevant literature suggests that there is a considerable degree of agreement in this respect. Nevertheless, an attempt to formulate a definition of the concept seems justified, at least from a theoretical point of view.

In all likelihood, the lack of a formal definition may be explained by two factors. Firstly, despite the growing number of academic works on the subject, scientific documentation and information is still, to a large degree, a practical field of study, with a very modest theoretical basis. There is a significant discrepancy between the interests and the theoretical background of the practitioners and that of the scholars. Secondly, the way of understanding a summary as a text which in a brief form conveys all the most important information (the meaning) of the work it summarises has caused problems. Any person with a rudimentary knowledge of semiotics perceives the difficulties related to understanding the message of a text. Moreover, the abovementioned intuitive description of the term 'summary'

[1]The author would like to thank Dr. Witold Marciszewski for reading and commenting on the first version the present work; owing to his remarks it was possible to supplement the current version of the article with comments on the possible objections to the (disputable) solution presented here.

does not simply pertain to any summary, but to a 'good' one. Thus, it adds new requirements for making an appropriate definition, namely the necessity to determine the criterion for comparing and evaluating various summaries of the same document. Such requirements seem justified from the practical point of view, and should therefore be taken into account in the process of formulating a proper definition.

The present attempt complies with the latter condition. It appears that — provided that the definitions proposed here prove acceptable — the problem can be solved by comparing the content of the documents without analysing the details about the matters discussed in the texts or determining the absolute scope of the content of a given document.

Understood as scientific information, documents are texts; thus, the present examination pertains to the set T, containing all texts. The elements of this set shall be labelled as t_i, t_j, t_k, etc. The set should also contain an element t_0, which shall be described as an empty text (i.e. a non-existing one). The necessity to include such a seemingly strange element was discussed elsewhere (Wojtasiewicz 1974).[2]

Each text may be described as a set of sentences; in this light, text t_0 is an empty set of sentences. The consideration does not have to be relativised to a single language, since the set includes texts in various languages; in practice, the summary of a given document is often written in a different language.

Furthermore, each text may be understood as a type, i.e. a class of tokens sharing the same form. This is the view implicitly adopted in practice, as it is entirely irrelevant which one of the 1000 copies of a given text shall become the subject of documentalists' work.

Let us consider the function D, which assigns a certain natural number to each text. The number is interpreted as the measure of the length of the text. We arrive at the following formula:

(1) $D\colon T \rightarrow N.$

Naturally:

(2) $D(t_i) \geqslant 0,$

[2]The concept of an empty text was introduced because legislative documents often contain references to specific regulations in executive decrees, etc., which may not yet exist at the time when the given legislative text is written — and are therefore empty texts.

(3) $(D(t_i) = 0) \Leftrightarrow (t_i = t_0)$.

(The length of any given text is not a negative number, and equals zero only in the case of an empty text).

The manner of calculating this measure is of secondary importance, yet some reservations must be made. Measuring the length of the text (i.e. the cardinality of the set) by the number of its sentences may initially seem the obvious choice (as mentioned above, a text is understood as a set of sentences). We would then arrive at the following formula:

(4) $D(t_i) = \bar{\bar{t}}_i$

It is, however, more advisable to measure the length of the text by different criteria, e.g. by the number of words it contains. An intuitive understanding of the term 'summary' implies that such a document should be shorter than the source text. The number of sentences would only be a valid criterion for comparison if the summary was composed solely of sentences taken from the source document. If even a single sentence is introduced to the summary by its author, the 'number of sentences' criterion would, at least from the theoretical point of view, prove unreliable: such a sentence could be so long that the summary would be longer than the actual text (if measured with criteria other than the number of sentences). A summary of this kind would probably be hard to find in real life, yet theoretical considerations ought to include precautions against such cases. Thus, a sentence (in a purely superficial understanding of the term) fails as a measure of the length of a text, at least for this particular task. Moreover, converting sentences (understood as surface structures) into some smaller semantic units, which would then be used as a criterion for evaluating the length of a text, would create more problems than it would solve.[3]

For this reason it might be more advisable to measure the length of a text on the basis of the number of words it is composed of. In the case of a written document (which is the form of most texts), a word should be understood as a sequence of letters preceded and followed by spaces or a non-letter symbol — with the possible exception of a hyphen, yet this issue is disputable and of secondary importance. This tentative definition may cause problems if, for instance, a Japanese document is summarised in the English language. In such cases the text written in a non-alphabetic

[3]A rather extreme example is Jerzy Andrzejewski's novella *Bramy raju*, which is formally composed of two sentences. The first stretches for several dozen pages, while the second only contains a few words.

script should first be translated into the language used for the summary (in alphabetic notation). In cases of comparing text in two non-alphabetic scripts (e.g. a Chinese document and a Japanese summary), both texts ought to be translated into a third language — one that uses alphabetic writing. In any case, such considerations are rather marginal in nature. From the theoretical point of view it might be assumed that it is always possible to compare the length of a source document with the length of its summary, performing certain auxiliary operations if need be.

Secondly, the measurement of the length of the text may, to a certain specified degree, be an approximation. This may also be considered a purely practical issue irrelevant for theoretical considerations. In this case the text could be measured with units, consisting e.g. of 10 words. The length of the text could then be presented as either the number of complete units or as the total number of units (including the incomplete ones). In the former case, the length measure '10' would apply to text consisting of 100—109 words; in the latter, to documents composed of 91—100 words. From a theoretical point of view, the matter may be of secondary importance — it could, however, have a degree of importance in practice.

To consider the issue further, we need to introduce the notion of a set of the closures of a given set of sentences, which shall be represented as $C(t_i)$ (i.e. the set if the closures of a given text); the notion of a set of logical tautologies, represented as L; and the notion of the set of all sentences, represented as S (Kotarbiński, Marciszewski, Czarnota 1970, entry: *Konsekwencja*).

It shall be assumed as obvious that the same text may have many differing summaries, including many different summaries of the same length.

The terms and premises specified above enables us to formulate a suggestion for a definition, with the proviso that it pertains only to a summary *sensu stricto*, excluding the bibliographical description of the summarised text (such a description is often a part of a summary *sensu largo*, sometimes referred to as an 'abstract'),

(5) $(t_j \in A(t_j)) \Leftrightarrow ((t_9 \neq t_0) \wedge \neg (t_j \subset L) \wedge (C(t_i) \neq S) \wedge (D(t_i)) \wedge (C(t_j) \subset C(t_i)))$.

The above definition should be interpreted as follows: t_j is a summary of t_i (or, to be more precise, an element of the set of summaries of t_i, i.e. one of the possible summaries of the text) if and only if: (a) t_j is not an empty text, (b) t_j is not composed exclusively of logical tautologies, (c) the set of closures of text t_i is not equal to the set of all sentences, (d) the length of t_j is less than that of t_i, (e) the set of closures of t_j is included in the set of closures of t_i.

The interpretation might need further explanation. From the practical perspective, requirements specified in (a), (b) and (c) may not only seem redundant, but even funny. They must, however, be included in order to eliminate eventualities which are theoretically possible, yet do not occur in real life (though precondition (c) may raise some doubts in this respect). A summary cannot be an empty text, because the set of closures of an empty text is included in the set of closures of all texts, and so, theoretically, if requirement (a) was not added to the formula, an empty text could serve as a summary to all existing documents (naturally, in practice an empty text is a non-existing one and, as such, cannot be treated as a summary).

Logical tautologies are closures of any given set of sentences — including an empty one. Thus, it is necessary to supplement the formula with requirement (b), stating that $t_j$cannot only comprise tautologies. Otherwise such a text would also serve as a summary for all documents.

Requirement (c) would not be fulfilled if $t_i$included contradictory sentences, since in this case the set of closures of text t_i would be equal to the set of all sentences, and would therefore incorporate the set of closures of every text. Be that as it may, many doubts arise at this point, since it is often impossible to determine whether a text that is being summarised complies with this requirement. There is no proof for the non-contradictoriness of arithmetics, and so all texts employing the apparatus of this branch of science are at risk of not being compliant with (c). It appears that the requirement should be included for the sake of theoretical accuracy, even though one has to bear in mind that it may often remain unfulfilled.

Requirements (d) and (e) pertain to matters significant from the practical point of view: (d) specifies that a summary must be shorter than the source texts, whereas (e) postulates that the content of a summary cannot exceed the content of the source document (it would then cease to be merely a summary, but would acquire the features of e.g. a commentary).

It is not necessary to supplement the *definiens* with a requirement specifying that $t_i$may not be an empty text. This prerequisite is implied in other elements of the definition and the previous assumptions: if t_j is not an empty text, its length must exceed 0; and given that the length of $t_i$is more than that of t_i, it must also exceed 0 and thus $t_j$is not an empty text.

Another obvious fact implied in the above definition is that no text can act as its own summary, because a summary must be shorter than the source document. Therefore:

(6) $\neg\ (t_i \in\ A(t_i))$.

Certain issues have been disregarded in formula (5): it is not relativised with regard to the reader of the text and their ability or capability to identify the closures of a given text. The reader is assumed to be if not ideal then at least possessing enough intellectual skill to arrive at the correct closures of a given text. This may certainly be viewed as a major simplification of the matter, yet in practice the users of documents are also assumed to have attained a certain level of intellectual competence. Furthermore, the definition could be expanded so that the last element of the *definiens* specifies that "for every user compliant with certain requirements the set of closures of $t_j$is incorporated in the set of closures of t_i."

The definition presented above contains one more controversial simplification. It disregards the fact that, in practice, the user arrives at its closure not only on the basis of a given text, but also by using their empirical knowledge relevant to a given issue. This constitutes a major complication, since it is often difficult to specify what knowledge may be considered relevant for a given issue, and many discoveries have been made precisely because someone noticed a connection where others had not. The final element of the *definiens* in (5) may be modified e.g. to:

(7) $(C(t_j \cup E) \subset C(t_i \cup E))$,

where E would signify a certain set of sentences representing empirical knowledge. Such an addition would not, however, constitute a valuable contribution for the purposes of the present analysis.

A more paradoxical consequence of definition (5) is that according to this formula each sentence taken from the text may be regarded as a summary of the said text, as it complies with all the requirements specified in (5). This is because (5) determines only the formal conditions that must be fulfilled for a given text to serve as a summary of a certain other text. Intuition dictates that a summary needs to be 'good', i.e. contain all elements included in the content of the text and considered 'important'. Thus, a definition of a 'good' summary would have to refer to a definition of an 'important' element of the content. Formulating such a definition appears to be a near impossible task, which would, at the very least, require extensive analysis. Perhaps somewhat surprisingly, it appears that it is possible to formulate a definition of an optimal summary of a given text with the proviso that only summaries of a certain specified length are taken into account. The suggested definition takes the following form:

(8) $(t_j \in A^n{}_{opt}(t_i)) \Leftrightarrow$

$$((t_j \in A)\ t_i)) \wedge$$
$$(D(t_j) = n) \wedge$$
$$\bigwedge_k((t_k \in A(t_i)) \wedge (D(t_k) \leqslant n)) \Rightarrow (C(t_k) \subset C(t_j))).$$

In this case the first requirement specifies that $t_j$is a summary of t_i, the second one determines that the length of the summary is n, and the third postulates that the set of closures of any summary of text t_i is included in the set of closures of t_i, if the length of such a summary does not exceed n. (Inclusion is understood in accordance with the tradition of set theory, as improper inclusion in the sense that every set is included in itself, i.e. constitutes its own subset.)

The above definition does not imply that for a given text there exists only one optimal summary the length of which does not exceed a certain specified n. Such an assumption seems intuitive: it is easy to imagine two summaries of the same document, both identical in length, but slightly different in content — i.e. formally differing summaries that would include the same content.

Definitions (5) and (8) appear to at least partially fill the gap in the theoretical representation of the term 'summary'. In spite of all the limitations of the method presented above, the possibility of comparing the extent of the content may be worthy of further consideration.

The notion of closure and a set of closures in research on natural languages was most probably introduced by Carnap. As regards Polish scholars, the terms were referred to by Irena Bellert (1972), but in a rather intuitive manner; Bellert appears to have disregarded the situations that might, in practice, be of little interest to a linguist. If the definition suggested above proves useful in theory (if in nothing more), it will be possible to employ the method in future research.

Bibliography

1. Bellert, Irena (1972) *On the Logico-Semantic Structure of Utterances.* Wrocław: Ossolineum.

2. Kotarbiński, Tadeusz, Marciszewski, Witold and Kazimierz Czarnota (1970) *Mała encyklopedia logiki.* Wrocław: Ossolineum.

3. Wojtasiewicz, Olgierd Adrian (1974) "Relacja odsyłania w tekście." *Zagadnienia informacji naukowej* 24[1]: 81-91.