On a set-theoretical model for cohesive and thematic structures of a text

Andrej Bekeš

0. Introduction

In this paper we are trying to develop a structure which could serve as a set-theoretical model for cohesive and thematic structures of a text, and examine some of its properties.

Object of our study will be ordered triplets (U, R, C) with U, R, C being some abstract sets. Triplets, where the three sets are related in a special way represent *information*, I.

In section 1. we shall introduce a binary relation in the set C, *linkedness*, which could serve as an abstract counterpart of cohesive relations. From this relation we develop the notion of *chains*, as subsets of C.

In section 2. we introduce the notion of *potential topic* as some subset of C, and the notion of a *potential text*, showing that each chain beginning in potential topic is a potential text. We further introduce the notion of a *theme* and show the relation between it and a potential text.

In section 3. we apply the results from the previous sections to the notion of *relevance*, a necessary condition connecting the theme and its predication in "senkoo" (apriori) type of thematization. We represent the relevance as an algorithmic process through which the theme and its predication can be connected. Most of the ideas used in this paper originally were developed by F. Daneš, P.

Sgall, M. A. K. Halliday, T. A. V. Dijk and appear here applied to simple abstract structures.

1. General background

1. 1 Definition: We take two sets, U (universum, a set of some abstract entities) and another such set, R (register).

Suppose that there is a mapping f from U to R such that:

a) for each $u \in U$ there is a $r \in R$, being picture $f(u)$ of u and at the same time, that for each r' there is a u' such that $r' = f(u')$;

b) that for x, y from U, such that $x \neq y$ it also follows that $f(x) \neq f(y)$. ■

a) and b) actually mean that f is a mapping from U onto R. and by definition we shall call such R a *register over U*.

1. 2 Definition: Take cartesian product $R^n = R \times R \ldots \times R$ (n − times) and suppose there is a subset C of R^n and that there is a mapping g from R^n to C.

Then we shall by definition call C a *context* and g a *restriction mapping* according to some grammar G. Also, we shall call the members of C *predications according to the grammar G* (or short, predications). ■

Predications have form $(r_1, \ldots, r_n)$, where r_i are *arguments*.

1. 3 Definition:

a) We shall call an ordered triplet (U, R, C) the *total information over universum U* and write it as I.

b) Let $U_1 \subset U$ and $R_1 \subset f(U_1)$ and $C_1 \subset g(R_1)$ be respective subsets as defined above. Then we shall by definition call the triplet (U_1, R_1, C_1) *information* I_1 and write: $I_1 \subset I$. ■

By the same procedure we can define an inclusion relation between two informations as well. Similarly, we define *union* and *intersection* between informations, as in the next definition:

1.4 Definition: We define union between informations $I_1 = (U_1, R_1, C_1)$ and $I_2 = (U_2, R_2, C_2)$ as $I_1 \cup I_2 = (U_1 \cup U_2, R_1 \cup R_2, C_1 \cup C_2)$; and intersection as $I_1 \cap I_2 = (U_1 \cap U_2, R_1 \cap R_2, C_1 \cap C_2)$; that is as respective operations between componets of the triplets. ■

1.5 Definition: We shall call two informations *unrelated* iff the intersection of their contexts is an empty set, i. e. iff for I_1 and I_2 we have $C_1 \cap C_2 = \phi$ ■

With this repertoire of concepts we can prove a simple theorem about unrelatedness.

1.6 Theorem: Let the informations $I_1 \subset I$ and $I_2 \subset I$ be such that $U_1 \cap U_2 = \phi$. Then a) I_1 and I_2 are unrelated, and b) from the fact that $C_1 \cap C_2 \neq \phi$ follows that $R_1 \cap R_2 \neq \phi$ and from this, that also $U_1 \cap U_2 \neq \phi$. ■

Proof: a) From the definition 1.1 follows that $f(U_1) \cap f(U_2) = \phi$. Otherwise there would exist an x as a member of this nonempty intersection. By 1.1 there would exist its original $f^{-1}(x)$ in the set U_1 as well as U_2. But this contradicts our assumption, that the intersection of U_1 and U_2 is empty and we must throw away this possibility. So, from $f(U_1) \cap f(U_2) = \phi$ follows that $R_1 \cap R_2 = \phi$ and from this $C_1 \cap C_2 = \phi$.

b) From the hypothesis of 1.6 b) we see that there is some P which is at the same time a member of both C_1 and C_2 and so there is a n-tuplet $(r_1, r_2, \ldots, r_n)$, member of both R_1 and R_2, and of course, their originals $u_1, u_2, \ldots, u_n$ are members of $U_1 \cap U_2$. QED

So we showed that if the two universums of respective informations have no common members, the contexts have no common members either, and also that informations, sharing a part of common context per force share also a part of common universe.

Let us now define another key concept, that of a *description.*

1.7 Definition: Let a be a member of R. Let us define the set $\bar{a}$ as the set of (all) predications P from C such that they contain a as their argument, i. e. all P of the form $P = (r_1, \ldots, r_i, a, r_{i+2}, \ldots, r_n)$ for some integer $i < n$. We shall by definition call such a set $\bar{a}$ the *description of a.* ■

This is a name figurative enough for a set of predications containing the same argument. Related to description is a concept of a *link* between two predications defined as below:

1.8 Definition: Let for P, Q from C there is an a from R such that P and Q belong to its description. Then a) we shal call such an a a *link between P and Q*; and b) say that such P and Q are *linked* (by a link a). ■

The concept of link is at this stage our abstraction of both thematic and cohesive relations within the text, which often appear superimposed in the texts as well. A further expansion of *link* is the concept of *chain* which we are going to define next.

1.9 Definition: Let $K = \langle P_1, P_2, \ldots, P_m \rangle$ be a m-tuple with $P_1, \ldots, P_m$ being members of C. We shall call such m-tuple a *chain* iff each two neighbours in it are *linked*, i. e. iff for each $i < m$, predications P_{i+1}, P_i are linked. ■

1.10 Corollary: Let a be a member of R. For each P, Q belonging to the description of a, i. e. $P, Q \in \bar{a}$, we can say that such P and Q form a chain. ■

Proof: follows immediately by applying 1.8 and 1.9. QED

1.11 Definition:

a) We shall call P, Q from C *related* iff there exists a chain K such that $\langle P=P_1, \ldots, P_m=Q \rangle$

b) Let k denote the number of links in the shortest chain connecting above P and Q. We shall write $k=d(P, Q)$, and call $d(P, Q)$ *distance* between P and Q.

c) For P and Q such that there is no chain connecting them, we shall say that $d(P, Q) = \infty$.

The most trivial example of P and Q related is when P and Q are linked.

2 . Text

2.0 Preliminaries

Here we begin supposing two entities, A and B called *participants*, each possessing information I_A and I_B respectively. For I_A and I_B we further suppose, that they are not unrelated.

2.1 Definition: Let T be a set of all $t \in U_A \cup U_B$ such that f(d) is a member of $R_A \cap R_B$. Then we shall call $\delta = g(f(T)) \cap (C_A \cap C_B)$ a *potential topic*. ■

A potential topic is a subset in C whose members can serve as we shall see later as centers around which the text will be spun.

2.2 Definition: A Quadruplet $<I, I_A, I_B, \delta>$ consisting of total information, informations belonging to participants A and B, and a set δ, fulfilling the condition in 2.1, we shall call *communicative situation*. ■

2.2 represents a static situation in which participants find themselves at any moment when their communication is interrupted. Next we shall try to introduce the time factor as well and make our model more dynamic. Before that, we make a convention, to denote each set from participants informations at a certain moment t_i with an upper index i, e. g. I_A^i is the information of A at the moment t_i etc.

2.3 Definition: Let $<I, I_A, I_B, \delta>$ be a communicative situation. If at the moment t_i, by either A's or B's producing of a predication P_i their respective informations are changed as follows we shall call such P_i a *minimal communication.*

a) $U_A^i = U_A^{i-1} \cup \{P_i\}$ and $U_B^i = U_B^{i-1} \cup \{P_i\}$

b) $R_A^i = R_A^{i-1} \cup f(P_i)$ and $R_B^i = R_B^{i-1} \cup f(P_i)$

c) $C_A^i = C_A^{i-1} \cup g(R_A^i)$ and $C_B^i = C_B^{i-1} \cup g(R_B^i)$

Here, P_i is a member of both C_A^i and C_B^i. ■

In the above definition we tried to express the fact that each produced predication enters participants information at all of the three different levels, i. e. his universum, his register and his context, and at the same time defined the triplet recurrently. Next, we shall give the notion of a *potential text.*

2.4 *Definition*: Let D $= \langle P_1, \ldots, P_m \rangle$ be an ordered m-tuple of members from C. If for each $i \leqslant m$ there is a chain $K_i = \langle Q_{1,i}, \ldots, Q_{mi,i} \rangle$ such that $Q_{1,i} = P_i$ and $R_{mi,i} \in \delta$, then, by definition we shall call such D a *potential text.* ■

Obviously, we can expand a potential text D so as to include chains stemming from δ (we shall call them δ chains) And this immediately gives us the following:

2.5 *Corollary*:

a) any δ chain is a potential text.

b) Let the chain K be such that for some P from δ, there is a Q from K such that P and Q are related. Then K is a potential text. ■

Proof: a) follows immediately from 2.4.

b) If Q is one of the ends of K it follows from 2.5 a). Otherwise we split K in two chains: $\langle P_1, ., Q \rangle$ and $\langle Q, \ldots, P_m \rangle$. Each of these chains forms together with the chain relating P and Q a δ chain and is accordingly a potential text. QED

2.1 Theme and information

Here we shall preoccupy ourselves with structures corresponding to thematical structures in real texts.

Let us for the sake of convenience denote: $I_A \cap I_B$ as I_{AB} and in the same way also for its componets, U, R, and C.

2.6 *Definition*: Let us define the following sets:

$C_A - C_{AB}$ = {all such P that P is member of C_A but not of C_{AB}}.

This represents A's exclusive knowledge of context C. And in the same way for B:

$C_B - C_{AB}$ = {all such P that P is member of C_B but not of C_{AB}}.

In the same way we define also $R_A - R_{AB}$, $U_A - U_{AB}$ etc. ■

Suppose now, that at the moment t_{i-1}, A's exclusive information, i. e. $I_A^{i-1} - I_{AB}^{i-1}$ is a nonempty set. The predications produced until this moment form an ordered set, $D^{i-1} = \langle P_1, \ldots, P_{i-1} \rangle$, and suppose also that A is the speaker. In this context, we shall say, that a production of a predication P_i, linking A's and B's common information I_{AB}^{i-1} with A's exclusive information $I_A^{i-1} - I_{AB}^{i-1}$ is a *communication* (from A to B).

The procedure of linking shall be formally executed in the following way:

Let P_j, $j \leqslant i\text{-}1$ a member of D^{i-1} be also a member of C_{AB} and let a be one of its arguments, i. e. P_j belongs to ā, the description of a. Let P_i also belong to ā. Then by 1.8 a is a link between P_j and P_i.

By 1.8, since P_j is a member of C_{AB}, a is a member of R_{AB} and by 2.1 a also belongs to potential topic δ.

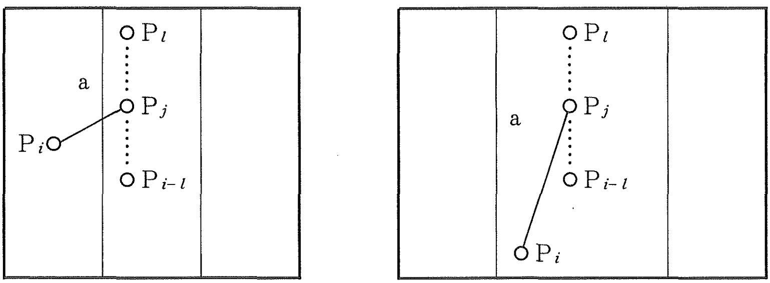
We suppose that each P_j, $j \leqslant i\text{-}1$ was produced in the same way as P_i. Then, the first predication P_1 had to be in C^1_{AB} at the moment t_1 and so also its arguments in R^1_{AB}. By 2.1 P_1 and any of its arguments form a potential topic and so we have proven the following theorem in the light of 2.6 the following.

2.7 *Theorem*: The sequence of predications linked with links whose descriptions at each step contain one of the preceding predications, belonging to C^{i-1}_{AB}, form a potential text. ■

2.8 *Definition*: links chosen as in 2.7 shall be called *themes*. ■

The above procedure also trivially satisfies the condition 2.3 for minimal communication, and the whole process of A passing P_i to B at the step t_i can be graphically represented as below.

$(C^{i-1}_A - C^{i-1}_{AB})$ C^{i-1}_{AB} $(C^{i-1}_B - C^{i-1}_{AB})$ $(C^i_A - C^i_{AB})$ C^i_{AB} $(C^i_B - C^i_{AB})$



t_{i-1} t_i

As a slight generalization of the type of text as it appeared in 2.8 we shall give the following definition:

2.9 Definition: let some potential text D be at the same time a δ chain $K = \langle P_1, \ldots, P_m \rangle$, where P_1 belongs to some potential topic δ. If.

a) for each $i \leqslant m$, $\langle P_1, \ldots, P_{i-1} \rangle$ is a subset of C^i_{AB} then, by definition, we shall call such D a *strong text.*

b) there is an $i \leqslant m$, such that for each $j : i \leqslant j \leqslant m$ the chain $\langle P_1, \ldots, P_{j-1} \rangle$ is a subset of C^i_{AB}, then, by definition, we shall call such D a *weak text.* ■

Before we state the next theorem about the existence of texts let us assume as an axiom, that A and B share in their registers the following elements: existential operator $\exists$ (i. e. "there is"), relation $\in$ (i. e. "being a member of some set"), and the concept of R_{AB} (i. e. "A and B both know the arguments.").

2.10 Theorem: If R_{AB} is a nonempty set, then C_{AB} is a nonempty set. ■

Proof: We have to show that there is at least one predication in C_{AB}. But by the above assumption, we have, for any x from R_{AB} at least the following predication, belonging to C_{AB}, i. e. $(\exists x \in R_{AB})$ — "there is an X that we both know." And this is also the basic agreement, necessary to begin with any kind of text. QED

Let us prove the following corollary about the possible choice of themes:

2.11 Corollary: Let x be an argument of some P, belonging to potential topic δ. Let also $(\bar{x} \cap C_A) - (\bar{x} \cap C_{AB})$ be a nonempty set. Then such x can be chosen as a theme. ■

Proof: from $P \in \delta$ follows that $P \in C_{AB}$ and also that $P \in \bar{x} \cap C_{AB}$. At the same time by $(\bar{x} \cap C_A) - (\bar{x} \cap C_{AB}) \neq \phi$ we have $\exists$ Q such that it is only $Q \in \bar{x} \cap C_A$. By substituting $P \rightarrow P_{i-1}$ and $Q \rightarrow P_j$ the condition from 2.8 is fulfilled, and so, x can be chosen as a theme. QED

2.12 *Theorem*: Let Q be some member from C. An argument from Q can be chosen as a theme iff Q is a predication in some potential text. ■

Proof: Let us first suppose that an argument from Q, for example x, can be chosen as a theme. Of course, $Q \in \bar{x}$. By 2.8 there is a δ chain K, such that there is a $P_i \in K$ and at the same time $P_i \in \bar{x}$. So P and Q are linked and $Q \in K'$ where $K' = \langle P_1, \ldots, P_i, Q \rangle$ is a δ chain and by 2.6 a potential text.

It remains to prove the reverse. Let us assume that Q belongs to some potential text D. By 2.4 we can expand D to a δ chain $K = \langle Q_1, .. Q_m, Q, Q_{m+1}, .., Q_n \rangle$, where $Q_1 \in \delta$ Applying the same linking procedure as in the proof of 2.7 to $(Q_1, Q_2), \ldots (Q_i, Q_{i+1})$ etc, until we reach (Q_m, Q) and at the $(m + 1)$ - st step, x from Q, such that both Q and Q_{m+1} belong to $\bar{x}$ is by 2.8 a theme.

3. An attempt to formalize the notion of the *relevance* condition between a theme and its predication.

3.0 "Senkoo" thematization

In Japanese as in many other languages (cf Mikami (1969), Dik (1978)) can be distinguished a pattern of thematisation which Mikami called "senkoo" or apriori, where the thematic element itself often appears to be in no structural relationship with the rest of the predication. In such cases, the only connection between the two is via the pragmatic relation of relevance. We shall represent such type of a sentence as:

x_{theme} $(x_1, \ldots, x_n)$ or $(x_{theme}$ P)

Some logicians have proposed "modus ponens," that is the logical relation of implication between the theme and its predication. But this turned out to be too restrictive, as the following formula tells us:

$(a \Rightarrow b) \quad \Leftrightarrow \quad (\neg b \Rightarrow \neg a)$

By it, if the comment of a sentence, such as the one below:

"Peter, he is ill."

were not true, then the existence of Peter himself would be false too.

Shibatani (1978) proposed the relation of inclusion, i. e. $X \supset x$ for some cases, but this is just a set-theoretical expression of the implication, so, by the above it is not suitable.

3. 1 An approach based on our model

Consider at the time t_i, with A as a speaker, the following predication with a "senkoo" type of the theme:

(x P) (with the condition that P is not a member of $\bar{x}$)

How can B in our model establish the relevance between x and P ? Let us proceed in steps.

Step I. x is a theme, so obviously, $\bar{x} \cap C_{AB}$ is nonempty. For the sake of convenience let us write

$\bar{x}_{AB}$ instead of $\bar{x} \cap C_{AB}$. Suppose, there is a Q, a member of $\bar{x}$, such that for some $y \in R_{AB}$ we have Q, $P \in \bar{y}$. Then, Q, with x as a theme and P form a chain, via y, thus also connecting together x and P. In the case there is no such Q, we proceed to

Step II. From $\bar{x}_{AB}$ we form $\bar{x}^2_{AB}$ defined as a set of chains with two elements, such that one of them belongs to $\bar{x}_{AB}$.

Suppose now, that there is a $Q \in \bar{x}^2_{AB}$, such that for some $y \in R_{AB}$, Q, $P \in \bar{y}$. Then, there is a $Q_1 \in \bar{x}_{AB}$ such that for some y_1 Q_1, $Q \in \bar{y}_1$. So we have x and P connected via two steps.

In the case there is no such Q in the second step either, we may continue the algorithm in the same way.

3.1 *Definition*: Let (x P) be a predication with a "senkoo" theme. If there is a k, $k < \infty$ such that P can be reached from x after k steps, we shal by definition call such x and P : k - distant. ■

Note: In the other type of thematization, where x is structurally a part of the predication, we can say that x and P are O (zero) distant.

Let us prove the following corollary:

3.2 *Corollary*: If for some predication (x P) with a "senkoo" theme there is a $y \in R$ such that $P \in \bar{y}$ and there is a k, $k < \infty$ such that $\bar{y} \subset \bar{x}^k_{AB}$ then x, P are k - distant. ■

Proof: Suppose there is such a y. Then from $P \in \bar{y}$ follows $P \in \bar{x}$ and from this by definition, the corollary. QED

With the above approach to the question of relevance in our model, we succeeded not only in establishing a relationship between the predication and its theme but also to introduce the degree of distance. This enables us to predict that the higher the k, the bigger the difficulty of a hearer to establish the relevance between the theme and the predication, and so the lower the acceptability of such a sentence in that particular context.

Explanation of signs and symbols: (in the order as they first appeared)

General:

U : "universe"
R : "register"
C : "context"
$P = (r_1, \ldots, r_n)$: "predication"
r_i : "argument of P"
f, g : "mappings"
f^{-1} : "inverse mapping"
G : "grammar"
i, j, k, l, m, n : "natural numbers"
(, ,...,) : "ordered set"
$<\ ,\ ,\ldots,\ >$: "ordered set"
I, I_i : "information"
$\bar{a}$: "description of a"
K : "chain"
d(P, Q) : "distance"
A, B : "participants"
t_i : "moment i"
I_A etc : "A's information" etc
C^i etc : "C etc at the moment i"
$D = <P_1, \ldots, P_m>$: "text"
δ : "potential topic"
$\bar{x}^k$: "description of x, of the order k"

$\blacksquare$: "end of the definition or theorem"
QED : "End of the proof"

Set theory:

ϕ : "empty set"
$a \in B$: "a is a member of B; a belongs to B"
$\neg$: "negation"
{ } : "denotes a set"
$A \cup B$: "union of sets A and B"
$A \cap B$: "intersection of sets A and B"
$P \Rightarrow Q$: "implication; from P follows Q"
$P \Leftrightarrow Q$: "equivalence; P equivalent to Q"
$A \subset B$: "inclusion; A is included in B"
A X B : "cartesian product of A and B"
A^n = A X A...X A (n times)
$\exists$: "exists; there is"
iff : "equivalent; if and only if"

Sources:

Daneš, F. (1966): "A three-level approach to syntax" Travaux linguistiques de Prague, 1966/I.

Dijk, TAV(1972): "Some aspects of textual grammars" Mouton, Hague 1972.

———(1977): "Text and context" Longman, 1977.

Dik, S. C. (1978): "Functional grammar" North - Holland, 1978.

Halliday, MAK and Hasan, Ruqaiya (1976): "Cohesion in English" Longman, 1976.

Mikami, Akira (1969): "Zoo wa hana ga nagai" Kuroshio shuppan, 1969.

Prochazka, O. and Sgall, P. (1976): "Semantic structure of a sentence and formal logic" Prague studies in mathematical linguistics 5, 1976.

Shibatani, M. (1978): "Nihongo no bunseki" Taisyuukan, 1978.

テキスト構造の集合論的モデル

ベケシュ・アンドレイ

この論文では、集合論によるテキストのモデルとなりうる構造を取り扱う。対象となっているのは（U, R, C）という順序のある triplet である。ここで集合 U, R, C はある抽象的集合で、triplet (U, R, C) を「情報」という。

第１節では、集合Cにおいて「継り（linked)」という２つの要素の間の関係を導入し、それによって「鎖」(chain）というこのモデルにおける基本的な概念を設定する。

第２節では、集合Cにおいて、「潜在的主題（potential topic）」と「潜在的テキスト（potential text)」の概念を設定する。そして、「潜在的主題」を起点とする鎖が、「潜在的テキスト」であることを示す。

更に、「題目 (theme)」を導入し、「題目」と「潜在的テキスト」の関係を明らかにする。

第３節では、今までの結果を、「先行題目」と叙述（predication）との「関連性（relevance)」の条件を限定するのに利用し、「関連性」の条件を algorithmic process として表す。