ISSN 1648-2824 KALBŲ STUDIJOS. 2002. NR. 3 * STUDIES ABOUT LANGUAGES. 2002. NO. 3

Some New Approaches to the Semantic Representation of the Text Content

Irina Liokumovich

Abstract. The paper deals with a new approach to the semantic representation of the content of scientific-technical texts in the field of civil engineering as viewed from the formal and quantitative perspective. It is based on the matrix and chain structures in conjunction with linear and paradigmatic verbal characteristics.
The semantic content of the scientific-technical text is realised by the object-process matrix, which shows an overall static picture of correlations between the semantic subclasses of key nouns with the semantic classes of verbal predicates. The matrix is seen as a semantic model of object-process coordinates on the level of categories and logic. Such matrices are developed for the text, group, corpus of texts under study. It enables to reveal patterns of text structure and the general semantic character of texts on civil engineering.
The dynamic aspect of the text content is represented by the linear "verbal-nominative chains" analysed on three levels: 1) level of lexics and semantics; 2) level of lexics and grammar; 3) level of object-process coordinates.
The study results have a practical application in automatic processing of the text and development of the system of the text synthesis based on matrix and linear structures. In addition, the study outcomes can be used in giving a course of lectures on textlinguistics.

Introduction

In recent years we witnessed a revival of interest in textlinguistic studies. Within current linguistic trends language structure and language functioning are no longer regarded as opposites but rather as complementary and interacting areas of linguistic research. The focus is on a coherent written text as a linguistic object with the aim of studying textual structure and procedures.

Scholars explored the role of various structural elements in the development of the semantic content of scientific-technical text (e.g. nouns, adjectives, etc.). It is generally acknowledged that a scientific-technical text has a nominative nature that is its main content is expressed by nouns. However, the issue of revealing the semantic content of the text, particularly the relationship between nouns and verbs in the process of its development has not been explored yet. This makes the study urgent.

The aim of the study is to find a new formal and quantitative approach to the semantic content of the text in the field of civil engineering.

To achieve this aim we should accomplish the following tasks:

1. To find out regularities of the linear development of the main content of the text on 3 levels of analysis: 1) level of lexics and semantics; 2) level of lexics and grammar; 3) level of object-process coordinates.

2. To find ways of formal representation of the semantic structure of the text, subgroup, group, overall corpus of text under study.

3. To single out the main tendencies of correlations between the semantic classes of verbs with the semantic subclasses of key nouns describing the main content of the text based on the object-process model.

Material of the study is a corpus of 45 English scientific-technical texts on civil engineering comprising 50,000 word usages selected on the basis of a common theme (e.g., buildings, bridges, tunnels).

The **object** of the study is various types of relationship between semantic subclasses of key nouns and semantic classes of verbal predicates.

The methods used are semantic modelling, transformational, contextual and statistical analysis.

Theoretical Background

The paper is based on the proposition that a written text refers to the linguistic object of the highest level. It reflects some real or mental situation that can be perceived from the author's viewpoint. A text is composed of passages and sentences that reflect some fragments of the total situation described in the text (Bell, 1991:166; Brown, 1966:157; Brown, Yule, 1983:97; Dijk van, 1997:231). In accordance with this approach we can single out two main constituents of the text:

- the static constituent;
- the dynamic constituent.

"Statics" and "dynamics" of the text can be seen as specific and contradictory features of the text. In linguistics the terms "statics" and "dynamics" have a multiple meaning. According to some scholars "statics" is defined as a state of balance while "dynamics" – as a state of motion. The text viewed as successive lexical elements as well as a product of the individual's mental and speech activity is in a static state (i.e., a state of balance). In this case the features of motion are implicit in the text. On the

contrary, the text in the process of the individual's production and comprehension is in a dynamic state (i.e., a state of motion). Therefore, the features of balance are implicit in the text (Galperin, 1981:19; Novikov, 1983:31).

Other authors consider two features of the text:

- vocabulary;
- syntax.

These parameters are interpreted differently by different authors (Longacre, 1996:73; Zubov, 1985:55).

We assume that for the current study the most appropriate viewpoint is that the term "vocabulary" means "content" while the term "syntax" implies grammatical structures of the text (*ibid*, 73, 55).

Thus, the term the "text content" ("vocabulary") is used to describe the static aspect of the text reflecting a great variety of real life objects. Since on the surface level of the text real life objects are expressed mainly by nominative lexical elements the formal structure of the text can be realised by linear successive key words – nouns. These studies are associated with making up of all kinds of graphs (Kamp and Reyle, 1993:64).

The term "syntax" implies the dynamic aspect of the text denoting the relationship between the objects of real life as perceived from the author's angle. In other words, the dynamic aspect of the text includes events (processes, actions, states) in which the object are involved. On the surface level of the text the text dynamics may be represented by a list of verbal predicates knitted by key words – nouns.

It is important to note that the text statics is closely connected with the text dynamics.

The Study

In this study the main content of the text from the static perspective may be expressed by key words – nouns that are subdivided into two main groups based on the quantitative criterion:

- main key words;
- secondary key words.

Lists of main and secondary key words are made keeping in view the absolute frequency of the word usage in the text and the number of passages in which the word occurs. Lists of supplementary key words expressing the optional content of the text are compiled based on the contextual analysis of the text.

As a result of the analysis of the study material we have made a semantic classification of key nouns on the paradigmatic level and semantic classification of verbs based on differential semantic components. On the whole, we have singled out 22 semantic subclasses of key nouns and 44 lexical and semantic groups of verbs that were distributed into seven lexical-semantic verbal classes.

Further, to study the principles of the development of the main content of the text we have made an attempt to single out linear verbal-nominative combinations – "verbal-nominative chains". These are made up of linear successive pairs "noun verbal predicate".

It should be noted that the study material has been extended by revealing implicit predicate relations by means of transformations. As a result implicit predicates, that are not found in the text or expressed on the level of type constructions (e.g., prepositional-nominal constructions, etc.) have become explicit on the sentence level.

N_1 prep. N_2, where N_1 is noun $_1$; prep. is preposition; N_2 is noun $_2$.

For example:
columns of the building $\rightarrow$ *T: The building has columns.*

It allowed the opportunity to describe a chain of actions in which each key noun is involved and give a set of its verbal characteristics. In other words, it enabled us to reveal the type and character of verbal-nominative chains, which were analysed on 3 levels:

- lexical and semantic level;
- lexical and grammatical level;
- level of object-process coordinates.

Each type of the verbal-nominative combinations revealed each aspect of the text dynamics but all together they were intended to give the overall picture of the semantic content of the text from the standpoint of dynamics.

In order to represent the semantic structure of the text corpus from the formal perspective we have developed a potential object-process matrix giving the overall picture of correlations between the semantic subclasses of key nouns with the semantic classes of verbal predicates.

The object-process matrix can be seen as a semantic model of object-process coordinates on the level of categories and logic. The matrix reflects the relationship between classes of words from the standpoint of the text content, particularly the category of nomination with the category of process. This means that the matrix lacks morphological and syntactic means of expression of the nominative and verbal elements of the text (See table 1).

Where,

ACT – verbs, predicates of action;
SPC – verbs – predicates of space;
ST – verbs – predicates of statics;
DN – verbs – predicates of dynamics;
RLT – relative verbs – predicates;
QLT – qualitative verbs – predicates;
INF – verbs – predicates of information;
MD – modal verbs.

As shown in table 1, the matrix is composed of 22 lines and 7 columns. At the beginning of lines there are codes of semantic subclasses of key nouns. For example:

No. 1 – construction project 1 (buildings).

On the top of columns there are codes of lexical and semantic verbal classes. For example:

1 – verbs of action (ACT).

Table 1. Potential object-process matrix of the text corpus in the field of civil engineering

No	Object Coordinates	Process Coordinates							
		ACT	SPC		PH	RLT	QLT	INF	MD
			ST	DN					
		1	2a	2b	3	4	5	6	7
1.	Construction project 1 (buildings)	C(1,1)	C(1,2a)	C(1,2b)	C(1,3)	C(1,4)	C(1,5)	C(1,6)	C(1,7)
2.	Construction project 2 (bridges)	C(2,1)	C(2,2a)	C(2,2b)	C(2,3)	C(2,4)	C(2,5)	C(2,6)	C(2,7)
3.	Construction project 3 (tunnels)	C(3,1)	C(3,2a)	C(3,2b)	C(3,3)	C(3,4)	C(3,5)	C(3,6)	C(3,7)
4.	Project part	C(4,1)	C(4,2a)	C(4,2b)	C(4,3)	C(4,4)	C(4,5)	C(4,6)	C(4,7)
5.	Component of the project part	C(5,1)	C(5,2a)	C(5,2b)	C(5,3)	C(5,4)	C(5,5)	C(5,6)	C(5,7)
6.	Mounting parts	C(6,1)	C(6,2a)	C(6,2b)	C(6,3)	C(6,4)	C(6,5)	C(6,6)	C(6,7)
7.	Building materials	C(7,1)	C(7,2a)	C(7,2b)	C(7,3)	C(7,4)	C(7,5)	C(7,6)	C(7,7)
8.	Coating, boarding	C(8,1)	C(8,2a)	C(8,2b)	C(8,3)	C(8,4)	C(8,5)	C(8,6)	C(8,7)
9.	Specifications	C(9,1)	C(9,2a)	C(9,2b)	C(9,3)	C(9,4)	C(9,5)	C(9,6)	C(9,7)
10.	Rocks	C(10,1)	C(10,2a)	C(10,2b)	C(10,3)	C(10,4)	C(10,5)	C(10,6)	C(10,7)
11.	Building machinery	C(11,1)	C(11,2a)	C(11,2b)	C(11,3)	C(11,4)	C(11,5)	C(11,6)	C(11,7)
12.	Machinery parts	C(12,1)	C(12,2a)	C(12,2b)	C(12,3)	C(12,4)	C(12,5)	C(12,6)	C(12,7)
13.	Totality, group	C(13,1)	C(13,2a)	C(13,2b)	C(13,3)	C(13,4)	C(13,5)	C(13,6)	C(13,7)
14.	Class, type	C(14,1)	C(14,2a)	C(14,2b)	C(14,3)	C(14,4)	C(14,5)	C(14,6)	C(14,7)
15.	System, structure	C(15,1)	C(15,2a)	C(15,2b)	C(15,3)	C(15,4)	C(15,5)	C(15,6)	C(15,7)
16.	Space orientation	C(16,1)	C(16,2a)	C(16,2b)	C(16,3)	C(16,4)	C(16,5)	C(16,6)	C(16,7)
17.	Name of the building company	C(17,1)	C(17,2a)	C(17,2b)	C(17,3)	C(17,4)	C(17,5)	C(17,6)	C(17,7)
18.	Professionals	C(18,1)	C(18,2a)	C(18,2b)	C(18,3)	C(18,4)	C(18,5)	C(18,6)	C(18,7)
19.	Results of mental activity	C(19,1)	C(19,2a)	C(19,2b)	C(19,3)	C(19,4)	C(19,5)	C(19,6)	C(19,7)
20.	Technological processes	C(20,1)	C(20,2a)	C(20,2b)	C(20,3)	C(20,4)	C(20,5)	C(20,6)	C(20,7)
21.	Physical phenomena	C(21,1)	C(21,2a)	C(21,2b)	C(21,3)	C(21,4)	C(21,5)	C(21,6)	C(21,7)
22.	Time indices	C(22,1)	C(22,2a)	C(22,2b)	C(22,3)	C(22,4)	C(22,5)	C(22,6)	C(22,7)

Each matrix cell is designated by two numbers in brackets used after the letter C ("coordination"): line number and column number separated by a comma. For example:

C(1,1), C(1,2a)…c(i,j),
where i = 1, 2, …, 22; j = 1, 2a, 2b, … 7.

Each column of the matrix can be seen as an interaction of the same lexical and semantic class of verbal predicates and a potential set of semantic subclasses of nouns. Every line of the matrix may be regarded as a combination of various lexical and semantic verbal classes with the same semantic subclass of key nouns. Any intersection of the matrix line with the column can be viewed as a kind of coordination of object and process categories. They determine the coordination and location of the object from the standpoint of the typical situation in which it is involved.

The potential matrix is an open system. The number of lines and columns can be increased if new lexical and semantic classes of verbs and semantic subclasses of key nouns are marked in the text.

Thus, the potential object-process matrix can be used to develop:

- analogous real matrices based on the text;
- real summary matrices based on the text subgroup (e.g., "buildings for libraries", "buildings for industrial works").
- real summary matrices based on the text group (e.g., "buildings", "bridges", "tunnels").
- real summary matrix based on the text corpus under study.

In terms of methodology the procedure of the development of real matrices based on the text subgroup were developed by summing statistical data of the matrices based on each text. Second, the matrices based on the text group were built by summing the data of the matrices based on each of three subgroups of text. Finally, the matrix based on the overall text corpus under study was developed by summing the data of the matrices based on each of three groups of texts.

Thus, real object-process matrices enable us to reveal general semantic image of the text, subgroup, group, text corpus under study. Additionally, the analysis of real matrices allowed an opportunity for us to single out three kinds of texts:

- texts of constructive-technological type (51.2%);
- texts of constructive type (24.4%);
- texts of technological type (24.4%).

Results

In practice the study results can be applied in automatic processing of the text and development of a formal procedure of the text synthesis based on the matrix and chain structures by making algorithms and programmes. In addition, the study outcomes can be used in giving a course of lectures on textlinguistics at higher education institutions.

Conclusions

To sum it up, a new formal and quantitative approach to the semantic representation of the content of scientific-technical texts in the area of civil engineering is based on the matrix and chain structures in conjunction with linear and paradigmatic verbal features.

We have made a formal and semantic decomposition of the text representing it by means of a series of real object-process matrices, which keep the syntagmatic features of the "dismembered" text. The matrices reveal major tendencies of correlation between the lexical-semantic verbal classes with the semantic subclasses of key nouns describing the main content of the text. As a result, it enabled us to reveal patterns of the text structure, a kind of general semantic image of the text, subgroup, group, overall corpus of texts under study from the static perspective.

The linear development of the main content of the text has been represented by means of "verbal-nominative" chains analysed on three levels:

- level of lexics and semantics;
- level of lexics and grammar;
- level of object-process coordinates.

It allowed the opportunity for us to reveal the main content of the text from the dynamic perspective.

References

1. Bell, R. T. (1991). Translation and Translating: Theory and Practice. London/New York: Longman.
2. Brown, G. (1996). Speakers, Listeners and Communication. Explorations in Discourse Analysis. Cambridge: Cambridge University Press.
3. Brown, G., Yule, G. (1983). Discourse Analysis. Cambridge: Cambridge University Press.
4. Dijk, T. A. van (ed.) (1997). Discourse Studies: 1. Discourse as Structure and Process. Hillsdale, New York: Erlbaum.
5. Kamp, H., Reyle, U. (1993). From Discourse to Logic: Introduction to Modeltheoretic Semantics of Natural Languages, Formal Logic and Discourse Representation Theory (Studies in Linguistics and Philosophy 42.). Dordrecht: Kluwer.
6. Longacre, R. E. (1996). The Grammar of Discourse (2nd ed.). New York: Plenum.
7. Гальперин, И. Р. (1981). Текст как объект лингвистического исследования. М.: Наука.
8. Зубов, А. В. (1985). Вероятностно-алгоритмическая модель порождения текста (семантико-синтаксический аспект). Дис....докт.филол.наук: 10.02.21. – М.
9. Новиков, А. И. (1983). Семантика текста и ее формализация. М.: Наука.

Irina Liokumovich

Nauji teksto turinio semantinio pateikimo metodai

Santrauka

Straipsnis pateikia naują požiūrį į semantinį mokslinių techninių tekstų turinio pateikimą statybos inžinerijos srityje iš formalios ir kiekybinės perspektyvos. Ji pagrįsta matricine ir grandinine forma kartu su linijinėmis ir tipinėmis žodinėmis charakteristikomis.
Mokslinio techninio teksto semantinis turinys suvokiamas kaip papildinio proceso matrica, kuri parodo bendrą statinį koreliacijų vaizdą tarp pagrindinio daiktavardžio semantinių poklasių ir veiksmažodžio tarinio semantinių klasių. Matrica suvokiama kaip semantinis papildinio proceso koordinačių modelis kategorijų ir logikos lygyje. Tokios matricos kuriamos tekstui, grupei, tekstų rinkiniui. Tai suteikia galimybę atskleisti teksto struktūros modelį ir bendrą statybos inžinerijos tekstų semantinį charakterį.
Teksto turinio dinaminė išraiška, pateikiama linijine "veiksmažodžio-vardininko grandine" analizuojama trimis lygiais: 1) leksikos ir semantikos lygis; 2) leksikos ir gramatikos lygis; 3) papildinio proceso koordinačių lygis.
Studijų rezultatai praktiškai taikomi automatiniam teksto apdirbimui ir teksto sintezei, kuri remiasi matricine bei linijine struktūra, ir sistemos plėtojimui. Be to, mokslo rezultatai gali būti panaudojami lingvistikos paskaitų kurso dėstymui.

Straipsnis įteiktas 2002 02
Parengtas spaudai 2002 11

The author

Irina Liokumovich, dr. assoc.prof. at Riga Technical University, Institute of Languages, Latvia.
Research interests: linguistics.
Address: Kalku str. 1, LV–1050 Riga, Latvia.
E-mail: irina.liokumovich@rtu.lv

DOI: 10.5755/j01.sal.1.3.43160