Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 31 (2014) 431 – 439

Procedia Computer Science

2nd International Conference on Information Technology and Quantitative Management, ITQM 2014

Specificities of Lexicological Synthesis of Text Documents

Boris V. Chernikov*, Aleksander M. Karminsky

National Research University Higher School of Economics, 20 Myasnitskaya Ulitsa, Moscow 101000, Russia

Abstract

In this paper, the characteristics of lexicological synthesis of slightly formalized text documents are presented. This technology provides a significant reduction in labour costs for the creation of text documents. It also improves text quality by reducing the probability of the occurrence of errors during the formation of documents and implementation of the requirements for their design features. An additional advantage of this way of creating documents is the decrease in the volume of stored information, as well as the improvement in the security of documents during their transmission over communication channels.

Selection and peer-review under responsibility of the Organizing Committee of ITQM 2014.

Keywords: lexicological synthesis; slightly formalized document; text document; reference word (term).

In any organization, a lot of managerial decisions on events and production processes are made on a daily basis. The more accurate the information about these events is, the more justified the decisions will be. Thus, it is extremely important to obtain timely and accurate information.

Information about various processes and events is usually transferred with the help of documents that accompany the production process or are created in connection with occurring events. Control functions are also expressed by relevant documents which may contain instructions, recommendations or requirements. Due to a constant increase in the complexity of business, crucial attention is devoted to the completeness and timeliness of the information used by management. Office functions that ensure the preparation of necessary documents are implemented in the form of documentation management that involves the processing of documents and the organization of effective work with them. The objectives of the latter are usually obtained by means of specialized automated systems. The aspects of the automation of creation, design and production of documents are underdeveloped and require special consideration.

Traditionally, management does not realize the benefits of implementing the new technologies during the automation of documents' preparation process. This happens because management often receives the final result in the form of prepared documents. A bevy of technological issues arise for executors who usually create documents by manually typing in the text using a keyboard. Unfortunately, this direct input method creates the possibility of making mistakes. Moreover, the following process of transfer either to other users, or to storage, carries the risks of insufficient protection during the transit. The significant size of files is yet another problem.

* Corresponding author. Tel.: +7-903-961-43-90
*E-mail address:*bor-cher@yandex.ru.

Selection and peer-review under responsibility of the Organizing Committee of ITQM 2014.
doi:10.1016/j.procs.2014.05.287

One of the peculiarities of modern text documents, which are used in systems of support of core activities and decision-making processes, is their weak formalization. This can be explained by the fact that many documents are formed in a "freely" style under insufficiently precise requirements for their contents. At the present stage of the process of documenting the information that supports the production processes, the following requirements should be followed:

- Data should be formalized in order to ensure the automated processing of information in the document;
- The creation of documents should take a minimum amount of time while maintaining the requirements for the information required to support decision making;
- Under the condition of weak formalization the prior formation of the content of the document into a form suitable for automatic generation of a specific document.

These requirements are consistent with both the Russian regulations and provisions of the European functional specification for managing electronic documents, i.e. European standard MoReq2. In addition to these requirements, MoReq2 pays a great deal of attention to the simplicity of usage and productivity that coincide with the requirements defined by the open systems' standard.

The traditional classification of the information in the documents implies permanent and variable information. Permanent information can be entered in documents' templates well in advance. The analysis of information from documents of various types used to organize different activities was conducted. The results illustrated that variable document information constitutes the largest part of a document, whereas the permanent information is contained in a small volume (Table 1).

Table 1.Information content of the document under traditional classification of information

Group of documents	Permanent information, %	Variable information, %
Organizational and administrative documents	13 – 18	82 – 87
Industrial enterprise documents	3 – 16	84 – 97
College documents	3 – 18	82 – 97
Hospital documents	6 – 18	82 – 94

The analysis of the information contained in the documents of various organizations illustrates that the proportion of permanent information is not high enough to produce a significant effect under significant variations in the content of documents. On average, the share of permanent information in the considered documents is:

- for organizational and administrative documents – 14.8%;
- for industrial enterprise documents – 8.1%;
- for college documents – 8.4%;
- for hospital documents – 8.7%.

Notice that the share values of permanent information are quite similar for organizational and administrative texts, but that for other types of documents there is much greater variety.

An analysis of the initial complexity of the creation of the documents, as well as their contents illustrates that the relatively low labour costs for the creation of organizational and administrative documents can be explained by a higher level of text typification. At the same time, the creation of documentation from other groups which have a lower level of typification, as illustrated by the small share of permanent information, is more time consuming.

It is possible to change this situation only by revising the traditional procedures of creating those documents. One may significantly improve the effectiveness of the process of creating documents through the use of lexicological synthesis. Synthesis technology (instead of direct text input) to create documents was used in various organizations for a wide range of activities. Approbation showed that lexicological synthesis is very effective in creating slightly formalized documents and the benefit level is proportional to the frequency of occurrence of the document: the more often the document is formed, the greater the benefit.

Slightly formalized documents – full-text, table-type, or mixed documents, the content of which significantly depends on an arbitrary structure, changing with the particular situation. These are the documents with a high degree of variability. In this regard, structuring of slightly formalized documents may require specification of interrelation

and mutual dependence of the composition of the text until the atomic values – fragments of phrases, words, and even parts of the individual words. According to the analysis, a significant number of documents accompanying the production process can be categorized as slightly formalized.

The principle of lexicological synthesis implies a preliminary analysis of a certain kind of documents, from which a set of sustainable phrases is extracted. Each of these wordings is assigned with a reference term. The choice of this term in the process of creating a document clearly defines the need for the introduction of specific wording. If the selected part of a text document contains a significant number of rows and is always presented in the document in a strictly defined sequence of sentences, this piece of text can be determined by a single reference term.

The traditional approach currently used in the typification of documents divides the information into the permanent and variable. However, it is more reasonable to use a deeper classification of information for automatic technology based on lexicological synthesis is. This method should take into account not just two (permanent and variable), but four categories of information (Fig. 1).

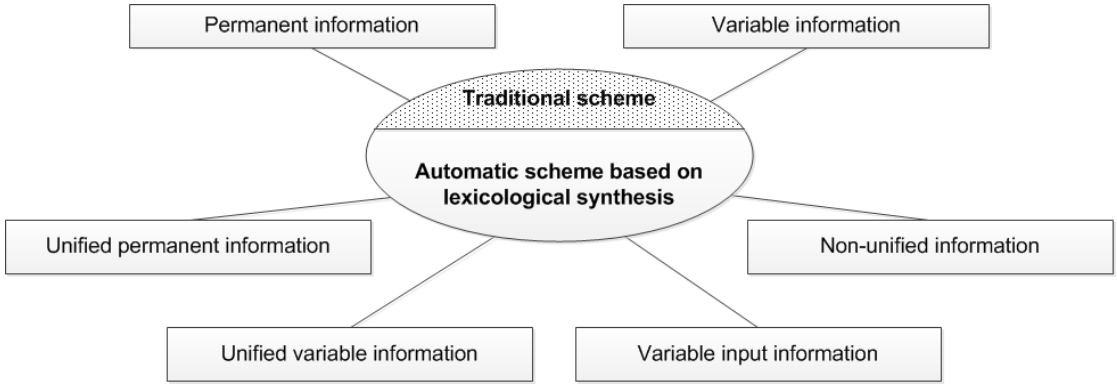

Fig. 1. Categories of information in the schemes of the documents' formation

In this case, it is reasonable to classify informational flows of documents in the following way:

- Unified permanent information that can be prepared in advance and stored in the database. It may also be contained in the program for the formation of the document. This information is automatically embedded in the created document. This includes permanent information (e.g., name of the document) and rarely changing information (names of subdivisions, personnel list, a list of sections of the document, etc.);
- Unified variable information. This type includes standardized and formalized data. This type of data is stored in the database and introduced in the document by selecting the necessary formulation. This type of information includes those formulations that are precisely offered to the contractor for selection during the synthesis of the document;
- Variable input information. It is subject to certain requirements for the presentation of data and includes specifying information. This information is used as a rule for a particular type of the document (e.g., table data, individual names, equipment specifications, data on the recommended work routine, assessment of monitoring activities, etc.). The information belonging to a given category is entered directly from the keyboard during the formation of the document;
- Non-unified information which contains free formulation and, if necessary, is introduced by typing directly from the keyboard.

The role of the documentation process in the improvement of document support in company management is great enough. That is why the studies were aimed at developing a methodology that would enable the user to create informal documentation provisions and create slightly formalized documents. Slightly formalized documents are used in management systems and are directly related to the conduct of the main activities of the organization. Such documents make up a very large share of all the company's paperwork.

Slightly formalized documents used in decision making aimed at observing events and facts are generated by automated lexicological synthesis. This is done by passing around the lexicological tree[1-3].

Each wording of the document is assigned with a basic term, the choice of which uniquely determines the specific wording in the document. These are reference words; they form the lexicological scheme of the document.

The set of mutually dependent reference words determines the bypass route for the formation of the document. Based on a preliminary analysis of the structure of the document, the main parts that will or may be present in the document are determined. The code names of such sections form the basis of the set of reference words. Within each section of the document, the main elements that should or may be included in the section are identified (words, phrases, or text fragments). For each element, the reference word (or set of words) is selected and later unambiguously determines the need to enter a corresponding fragment in the document. If a segment of the text document contains a significant number of rows and is always present in the document in a strict sequence of constructing sentences, this segment may be determined by a single reference word. In cases where the text of the document is generated from the sentences that are not fixed in a strict sequence, the number of reference words will be as large as necessary to unambiguously identify each individual sentence or phrase.

The full list of reference words including their interconnections forms the lexicological tree of the document. The 'passing' through the branches provides a choice of wording used in the document. The choice of certain reference words will signal the requirement to introduce an exact choice of text fragments into the document. In fact, the text of the document is formed by selecting the required number of storage formulations, each of which is mapped with a reference term. The structure of the lexicological tree is similar to the composition of the text document. The degree of branching depends on the number variations of the text document, defined by its complexity and the difference between documented situations.

In contrast to the 'direct' lexicological tree, in which the components are chosen in a mandatory sequence, the selection of reference words from the tree by a method of "pruning" is the more general occurrence. In this case, the choice of the next reference term depends on the one that was selected in the previous step. Moreover, for each specific copy of a particular type of document, the route of choice for the reference words is created by cutting off a number of lateral branches (Fig. 2, where thick line indicates a possible route for a set of reference words that make up the tree and which are selected during travel for the formation of the document). As a result, the lexicological tree formed using this methodology provides support for certain types of documents.

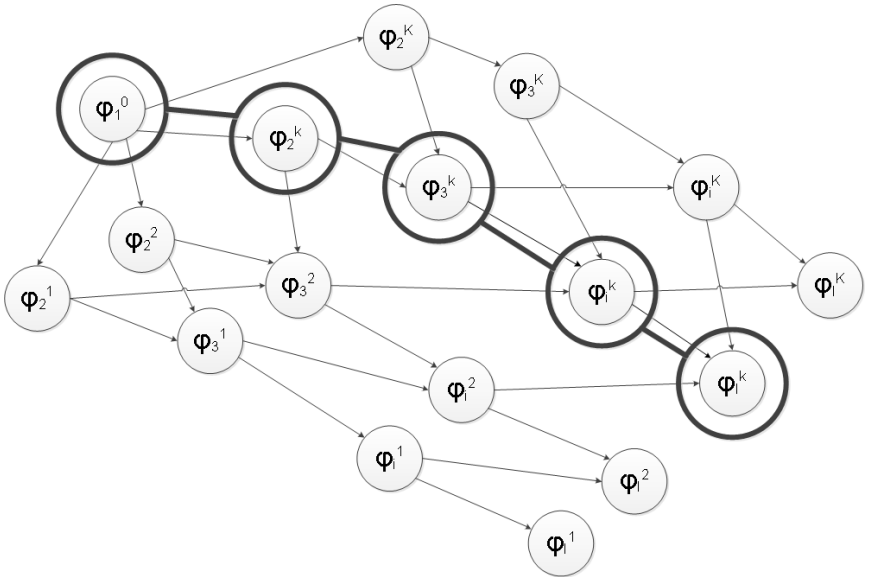

Fig. 2. Model of the formation of the document using lexicological tree with truncation

Reference terms may be different parts of speech that define the essence of the prescribed action.

A model of the formation of the document using a tree with truncation when branching of this type can be represented as follows:

$$D^{\mathrm{B}} \Rightarrow \sum_{i=1}^{I^{\mathrm{B}}} (\varphi_i \left| \varphi_{i-1}) \text{ при } \varphi_i \in \boldsymbol{\Psi}^{\mathrm{B}} \right. ,$$

where φ_i – current reference word ; I^B – number of reference words for the document D^B of concrete form; i – identification number (index) of the current reference word; $\boldsymbol{\Psi}^B$ – a set of reference words of this type of document.

By logical summation towards creation of concatenation of text document's fragments it is derived that not all reference words are selected, but only some of them, although all of them certainly belong to the set of reference words of this type of document.

When generating a lexicological scheme (which is a kind of ontological model of a document), as well as a lexicological tree, one should follow the principle of control of lexical structures and consider the relativity of the ontological status of reference words. Notions of ontological relativity have been expressed in the concept of linguistic frameworks by R. Carnap[4], who developed the idea of the multi-stage calculus of predicates. The reference word must be unique for a particular design, and if necessary it should be confirmed by other reference words, otherwise the choice of the desired fragment of texts can't be correctly identified. A clarification of one reference word by another forms their hierarchical subordination in the lexicological tree structure of a particular document.

The formation of lexicological schemes and the lexicological tree is based on the analysis of the relations of reference words, forming the route (path) of their choice during document formation. A lexicological scheme allows a relationship between reference words, considering that in different cases one can be observe a significant change in the route of documents' formation, which determines the possible variability of individual copies. An example of a lexicological scheme for the protocol of endoscopic inspection during gastroscopy is shown in Fig. 3.

At the stage of the automated document formation, (during the operation period) a database of formed documents' complex is used.

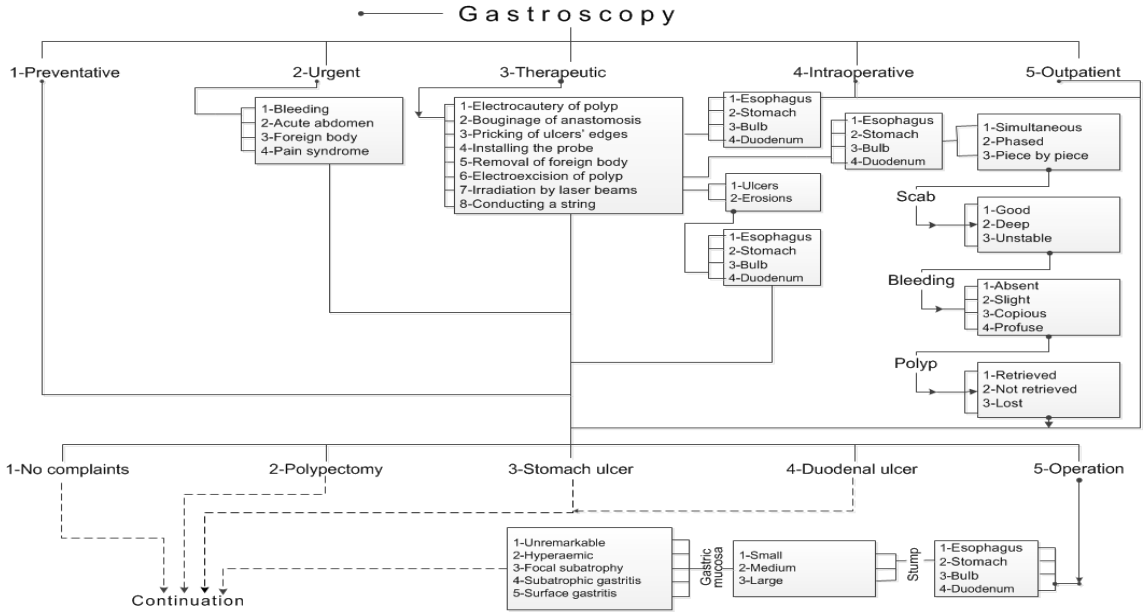

Fig. 3. Example of the lexicological scheme for a gastroscopy protocol

This database contains requisites, shapes, lexicological trees and sets of document reference words. When creating a particular copy of the document after selecting the first reference word the document fragment corresponding to the reference word is included in the text.

The process of selecting reference words is repeated throughout the document formation route considering the interrelation of reference words. Upon the completion of the selection of reference words, all the required details are introduced in a unified shape and a specific document if formed.

During the study, the model of automated technology for the creation of slightly formalized text documents is formed based on lexicological synthesis (Fig. 4).

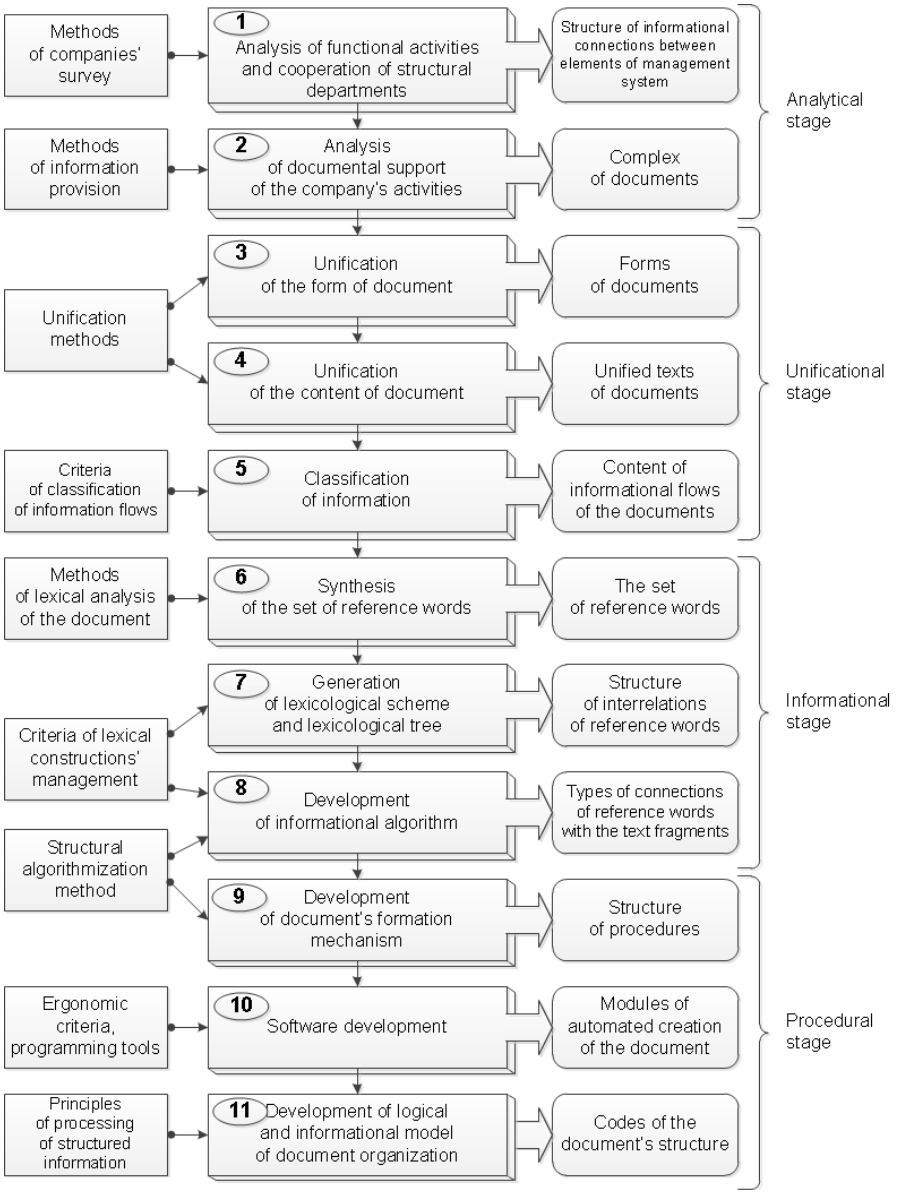

Fig. 4. Model of the development of technology of slightly formalized documents creation

The model structure is provided in four stages according to the tasks solved in each stage: the analytical, the unificational, the informational and the procedural. At the analytical stage, the complexity of a company's documents for which the technology of automated synthesis will be developed is listed. During the unification stage, the documents are processed in a similar form and content type. The importance of this stage is very great, since unification provides the possibility of sharing and reusing the designs of various documents. Moreover, here, the foundations of interoperability are created, i.e. the creation and processing of documents regardless of their technical and software platform.

It is at this stage that the prerequisites for time saving and lower material costs are formed. The informational stage of the development of lexicological synthesis of slightly formalized documents is used to build a specific info-logical data model. At the procedural stage, the preparation of the programs is made. It is then used for the direct implementation of the document-creation process, as well as to integrate the creation system with the informational system of the organization.

Lexicological synthesis of slightly formalized documents can significantly reduce the load on the storage system. When creating a document, the index sequence that is subject to further compression when creating the information package is generated. At the recovery phase the decoding of this package is made using the organized cycle of recovery based on the database of the lexicological tree[5]. The analysis of the effectiveness of the proposed method of processing of slightly formalized documents when organizing their storage shows the potential of significant (in the tens or even hundreds of times) reduction in size of stored information when compared with direct saving or compression of information. The graphical display of the size of stored documents for organizational, administrative industrial enterprise (Fig. 5), as well as college (university) and hospital documents (Fig. 6, where SEC is State Examination Committee) are presented.

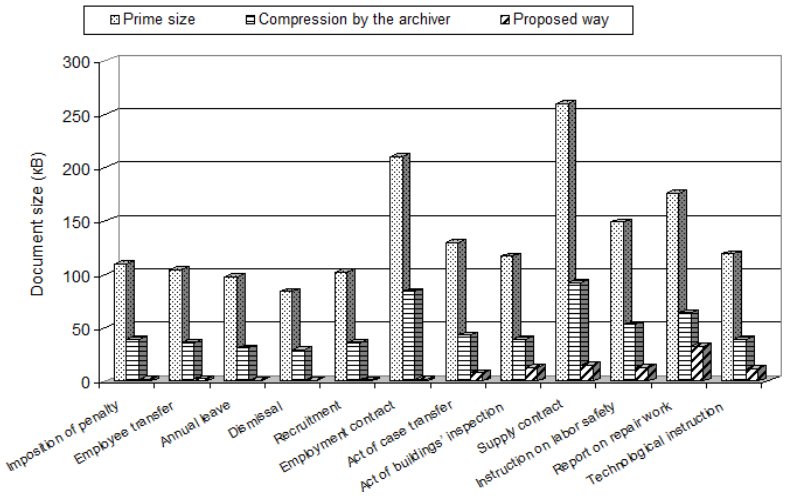

Fig. 5. Size of stored organizational and administrative documents and industrial enterprise documents

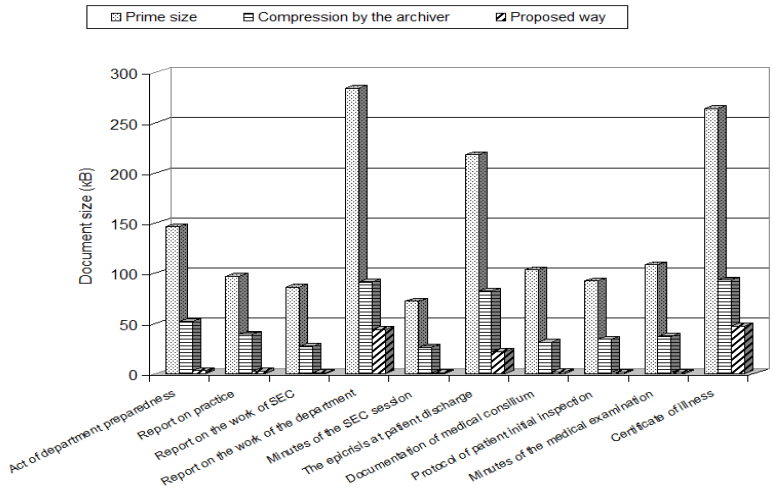

Fig. 6. Size of stored college (university) documents and hospital documents

Introduction of the automated method of lexicological synthesis to the practice of the formation of slightly formalized text documents moves group centers of accumulation of documents' components from the area of non-unified information towards unified elements. It predetermines the possibility of benefit under the automated choice of unified workings in the automated scheme (Fig. 7, where CD is College (university) Documents, OAD – Organizational and Administrative Documents, HD – Hospital Documents, DIE – Documents of Industrial Enterprise).

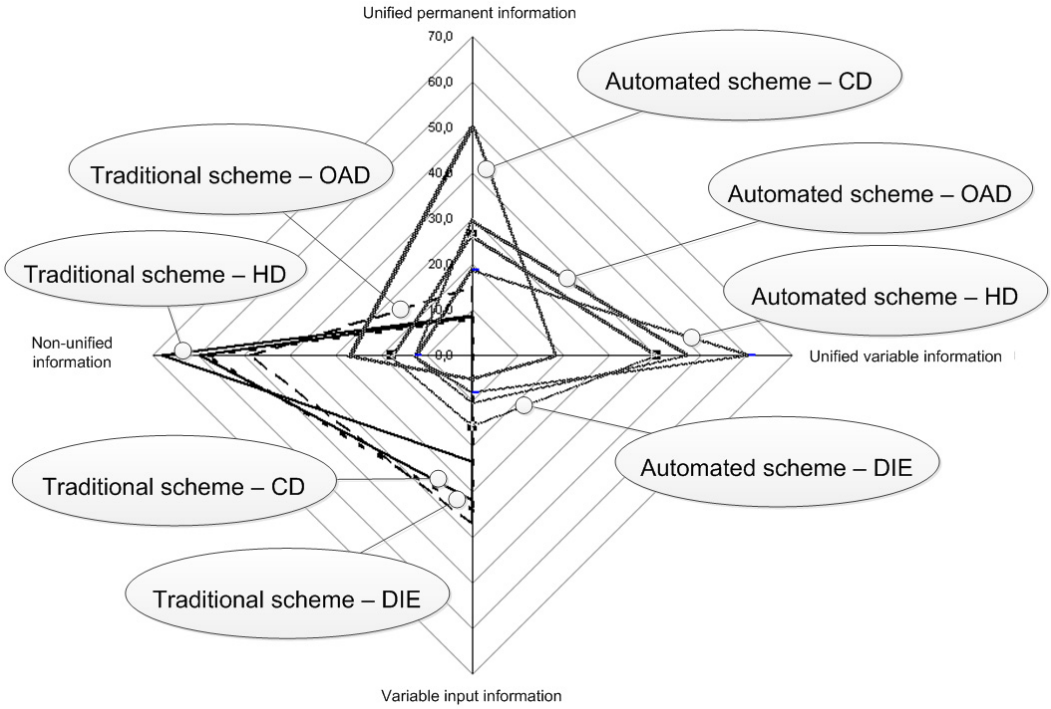

Fig. 7. Results of restructuring of information flows of documents to subsystems of documents

Note the significant growth of unified information in comparison to the traditional pattern of documents' formation (5 times the original in organizational and administrative documents, 8 times the original – in technical documentation and college (university) documents, 9 times the original – in hospital documents) and a significant decline in non-unified and inputted information (Table 2).

Table 2.Average rates of change in share of information components in the documents

Parameter	OAD	DIE	HD	CD
Growth of share of unified information	5.2	8.2	8.1	9.1
Reduction in share of variable input information	3.5	2.2	6.2	2.9
Reduction in share of non-unified information	3.9	3.2	2.2	5.4

The analysis of the labour content under the change of conditions of the documents' formation illustrates a decrease in labour costs amounting to double the original savings when the proposed technology is used to organize different types of business[6].

Graphically, the dynamics of labour content in the creation of organizational and administrative documents and documents belonging to the sphere of industrial enterprise during the implementation of lexicological synthesis is displayed in Fig. 8.

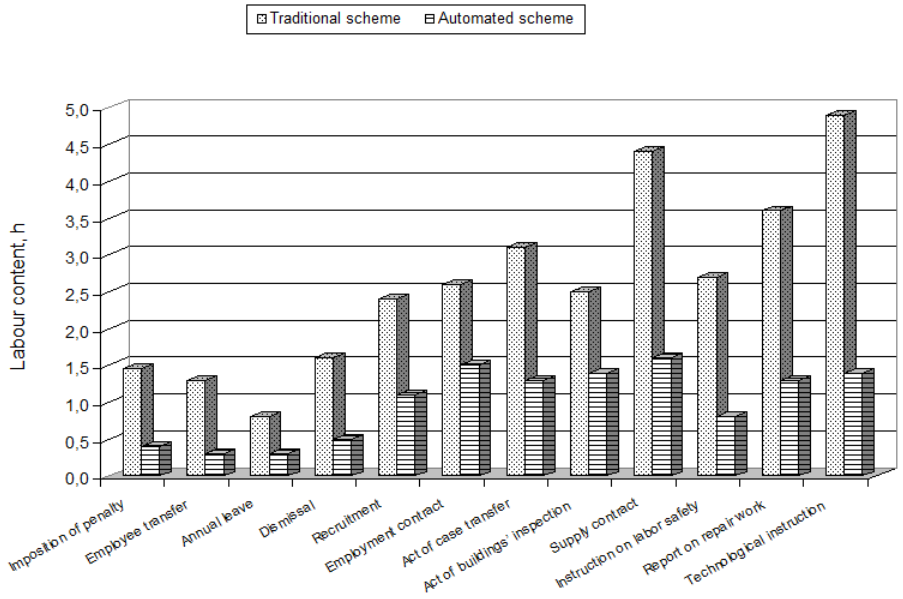

Fig. 8. The labor content of creating organizational and administrative documents and documents of an industrial enterprise

In general, the development and research of lexicological synthesis for text documents illustrated a high efficiency and a significant amount of potential for this technology. Additionally, it should be noted that the reduction of labour cost required for the creation of slightly formalized documents is accompanied by the improvement in the quality of generated documents.

References

1. Chernikov BV, Karminsky AM. The technology of automated formation of slightly formalized documents.In: Proceedings of III International conference "*Manegement of Large Systems Development*" MLSD'2009. Moscow, V.A. Trapeznikov Institute of Control Sciences of Russian Academy of Sciences; 2009, p.398-408(In Russian).
2. Chernikov BV. Lexicological synthesis of slightly formalized documents. Information Technologies and computing systems (Informacionnye texnologii i vychislitelnye sistemy). 2009. №4; p. 104-115 (In Russian).
3. Chernikov BV. A method of automated lexicological synthesis of documents. Patent of Russia № 2253893, 2005.
4. Carnap R. Meaning and Necessity. Studies in semantics and modal logic. Moscow: Oniks, 2012.
5. Chernikov BV. A method of converting of slightly formalized documents to minimize its size during storage. Patent of Russia №2413985, 2011.
6. Chernikov BV. Technology preparation of documents on the basis of cybernetic methods. Moscow: Finance and Statistics Publishing House (Finansy i statistika). 2009 (In Russian).