An Approach to Automated Thesaurus Construction Using Clusterization-Based Dictionary Analysis

Nadezhda Lagutina[†], Ilya Paramonov[†], Inna Vorontsova[*], Natalia Kasatkina[†]
[†]P.G. Demidov Yaroslavl State University, Yaroslavl, Russia
[*]Yaroslavl State Pedagogical University named after K.D. Ushinsky, Yaroslavl, Russia
lagutinans@gmail.com, ilya.paramonov@fruct.org, arinna1@yandex.ru, ninet75@mail.ru

***Abstract*—In the paper an automated approach for construction of the terminological thesaurus for a specific domain is proposed. It uses an explanatory dictionary as the initial text corpus and a controlled vocabulary related to the target lexicon to initiate extraction of the terms for the thesaurus. Subdivision of the terms into semantic clusters is based on the CLOPE clustering algorithm. The approach diminishes the cost of the thesaurus creation by involving the expert only once during the whole construction process, and only for analysis of a small subset of the initial dictionary. To validate the performance of the proposed approach the authors successfully constructed a thesaurus in the cardiology domain.**

I. Introduction

Thesaurus belongs to the most relevant tools for vocabulary control that applied linguistics and lexicography avail of. A terminological thesaurus reflects conceptual relations between terminological units and can be referred to as a model of the given domain. It can be used in systematic organization of scientific definitions and terms to show how they are related to one another as well as other terms belonging to the adjacent scientific areas. Thesaurus is irreplaceable for the efficient localization and definition of new terms, comparison and matching of these terms to their equivalents in the terminological systems of other languages [1].

The process of specialized thesaurus construction is labor-intensive and expensive as it requires processing of big amounts of text data and assessment of intermediate and final results by an expert [2]. In this regard, automation of thesaurus construction based on the use of information technologies seems to be reasonable.

In this paper we propose an approach for automated construction of a terminological thesaurus based on semi-automatic analysis of terminological dictionary corpora considered as arranged texts containing a set of terms and definitions that demonstrate semantic relations of hierarchical (genus/species) and non-hierarchical (synonymic, associative, etc.) types that can be used for divisions of these terms into semantic clusters.

As a reference case for thesaurus construction using the proposed approach we involve the cardiological lexicon. The choice of this area is determined by professional demand for the lexicographic resources of the kind since they are highly desirable for many categories of professionals including general practitioners, scientists, medical translators, medical students etc.

The paper is structured as follows. Section II contains a short introduction to the general principles of thesaurus creation and reveals motivation of our research. Section III overviews the papers related to automation of thesaurus creation. Section IV states the problem under consideration. The proposed approach is presented in Section V. The results of its application to construction of a terminological thesaurus for the cardiological lexicon are contained in Section VI. Conclusion summarizes main achievements of the paper.

II. General principles of thesaurus creation

A. Thesaurus definition and purpose

Thesaurus is a polysemantic term treated differently in philosophy, lexicography, information retrieval theory etc. In this paper we consider thesaurus as “a vocabulary of controlled indexing language, formally organized so that a priori relationships between concepts are made explicit” [1].

Such a thesaurus operates special lexical units—terms—that can be used for search (often for automated search) of documentary data. Each unit is matched with synonyms or correlated words standing in the semantic relations of broader and narrower terms, part and whole, means and objective etc. The existing variety of semantic relations is usually classified into genus/species and associative. Besides, a thesaurus may take a topic-based approach bringing together concepts of a specific domain.

The two main purposes of such a thesaurus include indexing of documents using concepts from semantic resources and enhancement of results of the user’s search request using hierarchical, associative, and synonymic relations. Additionally, the semantic relations between the words in the thesaurus can be used to classify and divide documents into clusters.

B. Thesaurus construction procedure

Thesaurus construction is based on creation of the domain model. The initial data for the modeling constitutes in a text corpus representative for the domain (dictionaries, textbooks, reference books, standards etc.). An expert constructing the thesaurus manually analyzes the corpus, makes a list of preferred index terms, and adds these terms to the thesaurus as keywords [3]. After that, he/she groups the terms into clusters and defines hierarchical and associative relations between them.

The macro-composition of the thesaurus is made up by a classification plan and a set of semantic clusters, e.g., “Heart anatomy (norm and pathology)", "Heart physiology (norm and pathology)", "Cardiac diseases (functional and organic)" etc.

The minimal unit of the thesaurus structure is an entry containing a detailed description of each term that belongs to a given cluster. The entry provides a user with information on grammatical categories of the term, data about grammatical and lexical collocability of the term presented in the form of patterns and illustrative examples. The entry of a thesaurus traditionally includes terms that stand in relations of semantic equivalence with the headword (synonyms and opposite terms) and related terms [4].

The main disadvantages of manual thesaurus-making are high cost, long duration, dependence of the result on expert's qualification, and restrictions of manual analysis of the large text corpus. Therefore approaches that automate the whole process or its stages are topical in modern linguistics.

III. Related Work

Computer-based lexicography has witnessed numerous attempts of automated thesaurus construction [5]. At the moment, it looks impossible to avoid participation of the expert in the process of thesaurus construction, however there are many works aimed at partial automation of either the whole process or its individual stages.

Most of the related papers consider thesaurus construction as a process containing several stages including extraction of a set of terms from the text corpus, distribution of these terms into clusters, and definition of the hypernym-hyponym relations between the terms.

For example, in [6] a semi-automatic construction of Jewish cross-period thesaurus is described. The authors propose an automated approach to extraction and grouping of terms into clusters based on absolute frequency of terms. In this process an expert is involved only for manual examination of the resulted set and removal of irrelevant terms.

It should be mentioned that the idea of using frequency as a main characteristic for term extraction is typical for most of papers. It is applied in many variations, e.g., in [7] the authors use the modified TD-IDF algorithm for that purpose.

The similar approach is described in [8] that is devoted to construction of the Chinese thesaurus using various electronic resources as corpora. In this paper the author does not use term clustering but define relation between the terms based on analysis of word combinations.

Terminological thesaurus can be considered as a domain model. In this regard, the process of its construction can utilize not only corpora in natural language but also semantically structured resources, such as dictionaries, encyclopediae, patents, technical or regulatory documents.

The paper [9] presents an automated system for general-purpose thesaurus construction based on an analysis of Portuguese dictionaries and texts. One of its key points is the using of synonymy relations retrieved from the dictionaries for term extraction. It should be mentioned that the synonymy being a particularly important concept for non-specialized thesauri is of less importance for specialized terminological thesauri mostly based on hypernym-hyponym relations.

The authors of [10] propose an automated approach for construction of the bilingual thesaurus based on analysis of Japanese and US patents. The hypernym-hyponym relations are extracted using the patterns (e.g., "A such as B"). Relations between the terms are established with the use of combination of machine translation and citation analysis methods.

Review of existing approaches to solve the problem of automation of thesaurus construction allows to identify the used methods and algorithms. It is worth mentioning that most of efforts in the existing works are concentrated on initial extraction of the terms and definition of relations between the terms, whereas automation of term clustering is studied insufficiently. Meanwhile, variety of existing clustering methods makes research in this area promising.

IV. Problem statement

As has been mentioned, in this paper we propose a solution for the problem of automated construction of the thesaurus for a specific domain.

As the initial data we use a monolingual explanatory dictionary in a specific domain. It is important to mention that the explanatory dictionary can cover a broader domain than the target one (e.g., for construction of the thesaurus for cardiology area a general medical dictionary would be appropriate). To start extraction of the related terms we also require a controlled vocabulary of 10–15 terms from the target domain provided by an expert.

The automation of thesaurus construction is aimed at the following operations:

1) Extraction of candidate terms for the thesaurus from the initial dictionary;
2) Distribution of the thesaurus terms into semantic clusters.

The other necessary operations are made by the expert.

The resulted thesaurus should have the following characteristics:

1) The set of all terms included in the thesaurus should be related to the target domain and should be reasonably complete (completeness is evaluated by an expert).
2) Each term should be accompanied by a description.
3) The thesaurus should have a hierarchical structure, i.e., should be divided into semantic clusters.
4) Between the terms (not necessarily between all of them) there should be established associative relations corresponding to the semantic relations of synonymy and antonymy.

V. Proposed approach

A. Overview

The proposed automated approach for thesaurus construction includes the following steps:

1) Preliminary processing of the text corpus (elementary text manipulations aimed at the selection of terms and

their descriptions to make them applicable for further automatic processing).
2) Automatic generation of a set of candidate terms, which can possibly be in the thesaurus.
3) Correction of the set resulted from the the previous step by the expert (the irrelevant terms are removed while the missing ones are added).
4) Automatic clustering of the terms into the clusters.
5) Estimation of clustering results and final grouping of the terms into the semantic clusters by the expert.
6) Establishment of the semantic (associative and hierarchical) relations between the terms of the resulted thesaurus by the expert.

The following subsections provide detailed explanations on each step of the approach.

B. Preliminary processing of the text corpus

Preliminary text processing is aimed at the extraction of all terms and their descriptions from the text corpus (an explanatory dictionary) represented as a text file. The dictionary entry consists of the so-called left part containing the headword expressed by a term or term combination, and the right part with the term's description. In the input file the left part ends with a period. The right part may begin with a stylistic, grammatical, or other label. The labels area is marked with square brackets. Since the labels do not contain any information required for thesaurus construction, they are removed during extraction of the right part of the dictionary entry. The term definition consists of a few sentences separated by periods. The words in the sentences are marked off with spaces, commas, and semicolons. The right part of the dictionary entry contains information needed for the construction of the thesaurus. Each dictionary entry finishes with a blank line.

We developed an automatic algorithm for preliminary processing. It determines the boundaries of the terms and definitions using a set of symbols that mark the end of elements described above, and extracts a term and its description from the dictionary entry. The term consists of one to five words while the description makes it up to 280 words. Upon the completion of this step we obtain a set of pairs $D = \{d_i = (t_{i1}, t_{i2})\}$, where t_{i1} is the left part of a dictionary entry and t_{i2} is the right.

C. Automatic generation of the candidate term list

The algorithm for automatic generation of the candidate term list developed by the authors takes as an input a controlled vocabulary of 10–15 key words and morphemes related the domain of the target thesaurus. The word search based on matching the full words as well as morphemes was expected to improve the quality of the resulted thesaurus.

The key word search in the dictionary entries was made using two criteria: by full words and by word morphemes. Using the second criterion relies on the fact that many terms belong to derived, hybrid, and compound words. Root morphemes with the highest frequency of occurrence were expected to enter into the morphological structure of many terms of the area considered, e.g., the terms "myocardium", "pericardium", "endocarditis" etc. contain the "-card-" morpheme.

It should be noted that the second criterion is only applicable for matching the left part of the dictionary entry because the latter contains a term expressed by a single word (accompanied by two or three synonyms in some cases) or a short word combination.

To select or reject a particular term we calculated two quantitative characteristics:

1) The number of key word (morpheme) occurrences in a dictionary entry (absolute frequency);
2) The percentage of key word (morpheme) occurrences in a dictionary entry (relative frequency).

For the estimation of results obtained at this stage by the expert, the dictionary entries selected were represented in two forms: as a list of terms sorted in the descending order of the quantitative characteristic, and as a list of terms placed in the alphabetical order.

D. Clustering

The purpose of text document clustering is to automatically identify groups of semantically similar texts among a given fixed set. In our case clustering allows to divide the candidate terms into semantic clusters.

It is important to notice that the text contains constructions which should not affect the clustering results: some general vocabulary, prepositions, articles, and numerals. It is easy to make sure that these types of words are automatically removed from all entries with the use of the preliminary formed set.

Besides, there are words that allows to directly put a term into a specific cluster. For example, the dictionary entries from the "The standard anatomy of heart and blood-vessels" contain the similar phrases: "between the atria of the heart" and "between the ventricles". In this case the occurrences of the words "atria", "heart", "ventricles" in conjunction with the word "between" can justify putting the corresponding term to the "The standard anatomy of heart and blood-vessels" cluster. However, the selection of such words requires lot of the expert's work because it is important to define particular clusters and corresponding relevant sets of terms in advance. In our research we are trying to minimize participation of the expert, and therefore do not use such an approach.

In our research we considered several algorithms for clustering of dictionary entries.

The k-means clustering [11] is based on the idea of referring each document to one of k predefined clusters. The clusterized objects are represented as vectors in the multidimensional space. Unfortunately, in our case it is difficult to represent a dictionary entry as a numeric vector. Moreover, preliminary determination of the number of clusters k implies the additional work for the expert. All these issues make such a method hardly applicable.

Another class of clustering algorithms that looked promising includes so known hierarchical algorithms [12]. The basis of the hierarchical clustering is the sequential fusion of clusters to form bigger structures, or the division of bigger clusters into smaller ones. The results of such clustering can be easily visualized. The initial data for the algorithm is the matrix of distances or similarities between the objects.

We have made an attempt of the hierarchical clustering to divide the dictionary entries into semantic clusters using the measure of similarity based on number of common coincident words in two dictionary entries. However, this approach did not give any results applicable for further usage because there was no conscious cluster division. The eventual reason for this was the inappropriate measure of the objects' similarity due to the fact that when a human composes a thesaurus his/her focus would not be on the common words in the dictionary entries but rather on terms related to a specific semantic cluster. It is rather hard to propose a numeric characteristics of such an expert's estimation.

There is an algorithm used for the text clustering named Suffix Tree Clustering (STC, [13]). However, it requires a considerable amount of preliminary text processing, such as the usage of a morphological analysis algorithm to change words according to the given grammatical form.

One of the algorithms that looks convenient for clustering of dictionary entries is called CLOPE [14]. It is intended for clustering of a set of transactions $T = \{t_1, t_2, \ldots, t_n\}$, where each transaction is a multi-set of items: $t_i = \{x_{i,1}, \ldots, x_{i,m_i}\}$. In our case transaction is a dictionary entry and items correspond to its words. Clustering $C = \{C_1, \ldots, C_k\}$ is a partition of T, i.e., $C_1 \cup \ldots \cup C_k = T$, $C_i \neq \varnothing$ and $C_i \cap C_j = \varnothing$ for all i, j such that $1 \leq i, j \leq k$, $i \neq j$.

The criterion function of a clustering C is evaluated as follows:

$$\text{Profit}_r(C) = \frac{\sum_{i=1}^{k} \frac{S(C_i)}{|D(C_i)|^r} \times |C_i|}{\sum_{i=1}^{k} |C_i|},$$

where $|C_i|$ is a number of transactions in the i-th cluster; $D(C_i) = \{x : \exists j \ x \in t_j \wedge t_j \in C_i\}$—a set of distinct items in the i-th cluster; $S(C_i) = \sum_{t_j \in C_i} |t_j|$—the total number of items in the i-th cluster; $r > 1$.

CLOPE finds a clustering C that maximizes $\text{Profit}_r(C)$ for given T and r. The r coefficient called repulsion allows to adjust possible similarity of transactions within a cluster and affects the number of resulted clusters: larger values of r imply more clusters generated by the algorithm. The specific value of r is selected by the user.

In this research the CLOPE algorithm showed the best results in clustering the dictionary entries and these results simplified the expert's work on cluster assessment.

VI. Evaluation

A. Candidate term list

The sampling of the dictionary was carried out from Online Stedman's Medical Dictionary[1]. It is designed to provide the language of medicine, pharmacy, nursing, and the allied health professions. It is made up of more than 100 000 terms and can be used to look up medical terms, abbreviations, acronyms, measurements, and more. Most definitions are also accompanied by pronunciation and word etymology.

The controlled vocabulary used to initiate the process of generation of the candidate term list was suggested by the expert (a practicing cardiologist) and contained fifteen key words and morphemes describing anatomical and physiological peculiarities of the heart and blood vessels including norm and pathology: heart, -card-, valv-, vessel, trunk, vascular, vein, artery, aorta, atrium, ventric-, block, hypertension, hypotension.

TABLE I. Examples of candidate terms selected as a result of the search over left parts of the dictionary entries

Key word/root morpheme (number of terms found)	Selected terms
heart (15)	heart; icing heart; beer-heart; heart-stroke; heart-failure; abrams heart reflex
card (105)	cardiohemothrombus, cardiothrombus; cardiopyloric; trichonocardiasis; stigmometric card; cardiopericarditis; mesocardium; pyopneumopericardium; endcarditis; embryocardia; electrocardiophonography; cardiophone; myocardium
valv(8)	valvulitis; valvotomy; valved; valveless
vessel (1)	blood-vessel
trunk (1)	trunk
vascular (7)	avascular; ideovascular; vascularization
vein (5)	veined; vein; hemorrhoidal vein;
artery (1)	artery
aorta (5)	aorta; aortarctia; aortic
ventric (20)	ventriculus; ventricle; intraventricular; ventricose; ventriculi quarti
block (4)	nerve-block; heart-block; blocking; block
fallot (2)	mefallotherapy; fallotomy
hypertension (0)	—
hypotension (0)	—
atrium (0)	—

Initially we performed the automatic key word search over the left parts of the dictionary entries. The search was based on root morphemes and used absolute frequency as the quantitative characteristic. As a result, we obtained 174 terminological units. Some examples of them are shown in Table I.

All the 174 terms met the requirement of belonging to the cardiology lexicon. It is an expected result because all the elements of the controlled vocabulary are either cardiological terms or parts of such terms. Unfortunately, the obtained set of words is far from being complete. For example, the following terms are missing: systole, diastole, (myocardial) ischemia, (myocardial) infarction. The resulted list is not enough to form a full thesaurus.

During the second phase of the candidate term list generation we considered the right part of the dictionary entries for the term selection. In this case the comparison method and the quantitative characteristic varied among the alternatives mentioned in Section V-C. In all cases the result was represented as a list of the selected terms sorted in descending order by the quantitative characteristic.

According to the expert's assessment, the most relevant results were achieved using the absolute frequency in combination with the root morpheme comparison criterion. Finally, as a result of automatic generation we obtained the list of 2 039 terms containing key words and morphemes. Table II shows several examples of such terms.

In the course of the results examination the expert notices

[1] http://stedmansonline.com/public/LearnMore.aspx?resourceID=Medical

TABLE II. EXAMPLES OF CANDIDATE TERMS SELECTED AS A RESULT OF THE SEARCH OVER RIGHT PARTS OF THE DICTIONARY ENTRIES

Term (number of occurrences)	**Description**
heart (37)	A hollow musciilar organ which receives the blood from the **vein**s and propels it into the arteries. It is divided by a musculo-membranous septum into two halves — right or venous and left or arterial, each of which consists of a receiving chamber (auricle or **atrium**) and an ejecting chamber (**ventric**le); the orifices through which the blood enters and leaves the **ventric**les are provided with **valv**es, the mitral and the aortic for the left **ventric**le, the tricuspid and the pulmonary for the right **ventric**le, armored h., calcareous deposits in the peri**card**ium occurring in subacute or chronic inflammation, bony h., the presence of more or less extensive calcareous patches in the peri**card**ium and walls of the **heart**, fatty h., (i) fatty degeneration of the myo**card**ium; (2) an overaccumulation of adipose tissue on the external surface of the **heart** with sometimes an infiltration of fat between the muscle bundles of the **heart** wall;. cor adiposum. fibroid h., chronic inflammation of the myo**card**ium, with overgrowth of the connective tissue. hairy h., peri**card**itis in which the **heart** is seen post mortem to be covered with a shaggy, fibrinous exudate; cor hirsutum, cor tomentosum, tricho**card**ia, shaggy peri**card**ium, icing h., peri**card**itis in which the **heart** is seen post mortem covered with a thick, white coat like the icing of cake, irritable h., soldier's **heart**, D.A.H., neurocirculatory asthenia, a **card**iac neurosis due to overstrain; marked by rapid pulse, dyspnea, and various neurotic symptoms, associated with an increased susceptibility to fatigue, observed especially in soldiers in active war service but noted occasionally also in civil life, left h., systemic h. lux'us h., a German term for combined dilatation and hypertrophy of the **heart**, of the left **ventric**le chiefly, pulmonary h., the right auricle (**atrium**) and**ventric**le, receiving the venous blood and propelling it to the lungs, right h., pulmonary h. skin h., the peripheral blood-**vessel**s, soldier's h., irritable h. systemlc h., the left auricle (**atrium**) and **ventric**le, receiving the aerated blood from the lungs and propelling it throughout the body, tiger h., a fatty degenerated **heart** in which the fat is disposed in the form of broken stripes. tobacco h., **card**iac irritability marked by irregular action, palpitation, and sometimes pain, occurring as a result of the excessive use of tobacco, absence, a**card**ia. atrophy, a**card**iatrophia, **card**latrophy, atrophia cordis, calculus, **card**ioHth. causing contraction, **card**iooinetic, **card**iokinetic. clotin.
aorta (8)	The main tnmk of the systemic arterial system, arising from the base of the left **ventric**le; the thoracic **aorta** is divided into the ascending portion, the arch, and the descending portion; at the diaphragm it becomes the abdominal **aorta** and bifurcates at the left side of the body of the fourth lumbar vertebra into the right and left common iliac arteries, a. abdominalis, the terminal portion of the **aorta**, extending from the diaphragm to the bifurcation into the common iliac arteries; its branches are the paired inferior phrenic, lumbar, common iliac, suprarenal, renal, and spermatic or ovarian, and the single middle sacral, celiac, superior and inferior mesenteric arteries, a. angusta, congenital narrowness of **aorta**, a. chlorofica, a general narrowing of the **aorta** associated with certain cases of chlorosis. a. thoracalis, thoracic **aorta**, the **aorta** from its origin to the diaphragm; its branches are the coronary, innominate, left subclavian and common carotid, intercostal, subcostal, diaphrag- matic, vas aberrans, bronchial, esophageal, peri- cardial, and mediastinal arteries.
infarction (2)	An area of coagulation necrosis resulting from the arrest of circulation in the **artery** supplying the part. anemic i., pale i. calcareous i., a deposit of calcium salts in the connective tissue, hemorrhagic i., red i. pale i., a whitish, bloodless area of necrosis caused by arrest of circulation in the terminal **artery**, or resulting from decolorization of a hemorrhagic i. red i., an area, red in color and swollen, the seat of hemorrhagic infiltration.

that the value of the quantitative characteristic did not correlate to relevance of selected terms, since even among the dictionary entries with a low values of the quantitative characteristic there were terms that should be added to the thesaurus (e.g., "infarction" in Table II).

The following examination of the terms' semantics showed that about 50 % of the automatically selected terms related not to the cardiological area, but to other medical fields, such as gastroenterology, orthopaedics, neurology etc. The high percentage of redundant terms can be explained due to the existence of homonym-terms (e.g., atrium, ventricle) and common usage of morphemes typical to cardiology also to noncardiological terms (e.g., hyper-/hypotension).

After filtering the resulted list by the expert we obtained a list containing 1 009 terms related to the cardiology lexicon.

B. Clustering results

In order to assess necessity of involving the expert before the clustering stage we made the clustering twice: using 2 039 terms from the automatically generated candidate term list and 1 009 terms filtered by the expert after examination of this list. In both cases the clustering was performed with the use of the CLOPE algorithm. The repulsion r varied in the interval $1 \leq r \leq 2$ with the step of 0.0001.

The expert estimated the results in which the number of clusters varied from 5 to 20 (it corresponds to a presumable amount of semantic clusters). It was achieved when the repulsion was in the interval $1.18 < r < 1.29$. Table III illustrates distribution of terms into clusters for different values of r.

According to the expert's conclusion, the results of clustering can be used as the initial working materials for inner structuring of the cardiological lexicon terms. Let us discuss them in more detail.

All clusters with less than 10 units contained the terms stood in the absolute synonymic relations or the terms being orthographic or morphological variations of a single term, such as dexiocardia/dextrocardia, pericardium/heart-sac, sphygmocardioscope/sphymocardiograph, bradycardia/brachycardia, aerocardia/araiocardia, and so on. These terminological combinations would be located in the final semantic clusters of the thesaurus that contain terms of the same generalization level.

Larger clusters (having from fifty to a hundred terms) contained terms related to different subfields of cardiology, therefore they required some additional manual processing. However, the expert's work in this case was rather easy because the the terms from the same area were usually followed in the resulted list one after another.

For example, it turned out that in the beginning of the fifth cluster for $r = 1.19$ there were terms related to the semantic cluster of "The standard anatomy of heart and blood-vessels": interatrial (between the atria of the heart), interventricular (between the ventricles), intravascular (within the blood-vessels or lymphatics), intra-atrial (within one of the atria of the heart), endocardium (innermost tunic of the heart), intravenous (within a vein or veins), intramyocardial (within the myocardium or wall of the heart), and periatrial (surrounding the atrium, or auricle, of the heart). These terms were followed by the terms related to the "Patology of heart and blood-vessels" cluster: phlebocholosis (disease of a vein), phlebectasia (dilatation of the veins, varicosity), vasoconstriction (narrowing of the blood-vessels), cardiopalmus (palpitation of the heart), cardiomegaly (hypertrophy of the heart), and capillarectasia (dilatation of the capillary blood-vessels). The following four units were from the semantic clusters of "The standard anatomy of heart and blood-vessels"—angiogenesis and intra-auricular—and "The manipulations"—phleborrhaphy and venesuture, but after those terms the list of "Patology of heart and blood-vessels" terms continued: telangiitis (inflammation of the capillary blood-vessels), cardiodynia (pain in the heart), angiocarditis (inflammation of the heart and blood-vessels), and omphalophlebitis (inflammation of the umbilical veins).

TABLE III. QUANTITY OF TERMS IN CLUSTERS CORRESPONDING TO DIFFERENT VALUES OF REPULSION r

r	No. of a cluster / Quantity of terms in the cluster																			
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**
1.18	1 508	391	114	16	4	1	1	1	1	1	1									
1.22	1 328	458	11	89	97	1	1	40	4	2	2	2	1	1	1					
1.28	1 126	58	2	1	3	54	1	1	22	6	116	1	479	75	81	4	3	4	1	1

We mentioned that the morphological proximity of terms increases performance of the proposed approach (e.g., common affixes such as "inter-", "intra-", "endo-", "peri-", "-itis", "-osis", "-ectasia", etc.). It is difficult to say whether the clustering method would be as efficient if the terms were not so close morphologically. However, recurring morphological elements are typical for many specific lexicons, making the proposed approach promising.

Here are the semantic clusters formed as a result of the clustering process:

1) Standard anatomy of heart and blood vessels;
2) Standard physiology of heart and blood vessels;
3) Pathology of heart and blood vessels;
4) Tools and instruments;
5) Pharmacology;
6) Surgical intervention and manipulations.

Clustering of filtered and non-filtered candidate term lists gave similar results, therefore it is possible to involve the expert only once during the whole thesaurus creation process: after the clustering many relevant terms occupy close positions in the clusters as well as irrelevant words, making elimination of irrelevant words simpler.

VII. Conclusion

In this paper we presented an approach for automated construction of the terminological thesaurus. It is based on the algorithm for selection of dictionary entries by key words (morphemes) and the clustering algorithm. The first one is used for creation of the thesaurus's corpus and the second one allows to form semantic clusters. Performance of the approach was validated by construction of the thesaurus for cardiological lexicon.

The examination of the results by the expert showed that the formed set of terms adequately represents the described domain. It should be mentioned that even the most complete dictionary does not contain the newest terminology, which has to be additionally extracted out of textual documents. However, the appeal to a verified dictionary makes the efficient base for terminological thesaurus construction.

In our reference case the clustering results were convenient for making hierarchical relations in the thesaurus. The words in a separate cluster were either related directly to one cluster or were easily subdivided into groups related to different clusters. Such an effect is presumably due to the morphological closeness of the terms inside the target domain. This should make the proposed approach applicable for construction of thesauri in other professional areas.

The future proposals include further research on the applicability of the proposed approach and dependence of the criterion function in the clustering algorithm on the quality of results.

Acknowledgment

The article was published with financial support by the project No. 549 of P.G. Demidov Yaroslavl State University within State Assignment for Research.

References

[1] J. Aitchison, A. Gilchrist, and D. Bawden, *Thesaurus construction and use: a practical manual*. Psychology Press, 2000.

[2] M. L. Nielsen, "Thesaurus construction: Key issues and selected readings," *Cataloging & classification quarterly*, vol. 37, no. 3-4, pp. 57–74, 2004.

[3] National Information Standards Organization (US) and others, *Guidelines for the construction, format, and management of monolingual controlled vocabularies*. NISO Press, 2005.

[4] L. M. Garshol, "Metadata? Thesauri? Taxonomies? Topic maps! Making sense of it all," *Journal of information science*, vol. 30, no. 4, pp. 378–391, 2004.

[5] G. Grefenstette, *Explorations in automatic thesaurus discovery*. Springer Science & Business Media, 1994.

[6] C. Liebeskind, I. Dagan, and J. Schler, "Semi-automatic construction of cross-period thesaurus," in *Proceedings of the 7th Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*. Association for Computational Linguistics, 2013, pp. 29–35.

[7] M. Zhang, S. Qu, T. Du, and Q. WANG, "Subject thesaurus automatic construction based on multidomain distribution entropy," *Journal of Computational Information Systems*, vol. 9, no. 9, pp. 3485–3492, 2013.

[8] Z. Wen, "Exploration and study of chinese thesaurus automation construction for digital libraries," *Journal of Convergence Information Technology*, vol. 6, no. 4, 2011.

[9] H. G. Oliveira and P. Gomes, "ECO and Onto. PT: A flexible approach for creating a Portuguese wordnet automatically," *Language resources and evaluation*, vol. 48, no. 2, pp. 373–393, 2014.

[10] H. Nanba, S. Mayumi, and T. Takezawa, "Automatic construction of a bilingual thesaurus using citation analysis," in *Proceedings of the 4th workshop on Patent information retrieval*. ACM, 2011, pp. 25–30.

[11] J. MacQueen, "Some methods for classification and analysis of multivariate observations," in *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, vol. 1, no. 14, 1967, pp. 281–297.

[12] F. Murtagh and P. Contreras, "Algorithms for hierarchical clustering: an overview," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 2, no. 1, pp. 86–97, 2012.

[13] O. Zamir and O. Etzioni, "Web document clustering: A feasibility demonstration," in *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, 1998, pp. 46–54.

[14] Y. Yang, X. Guan, and J. You, "CLOPE: a fast and effective clustering algorithm for transactional data," in *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2002, pp. 682–687.