Daria Ryzhova

AUTOMATING RESEARCH IN LEXICAL TYPOLOGY: METHODS AND TOOLS

Summary
of the thesis for the degree
of PhD in Philology and Linguistics

Moscow 2018

GENERAL OVERVIEW OF THE THESIS

Lexical typology is a field of linguistics that applies comparative analysis to word meanings across languages. Lexical typology received a powerful impetus with publication of the seminal work by Berlin, Kay 1969 on typology of colour terms; the authors introduced an elaborate methodology for comparing lexical meanings, which is still widely used today. Their paper had deep resonance with the linguistic community and spurred rapid developments in lexical typology. Presently, the interest to typological analysis of lexical meanings enjoys steady growth.

In the five decades of its development, lexical typology has accomplished remarkable achievements: a wide array of methods for data collection and analysis have been designed (e.g. see the recent overview in Koptjevskaja-Tamm et al. 2016), and a large number of lexical fields have been described (Newman 1998, 2009, Majid, Bowerman 2007, Майсак, Рахилина 2007, Kopecka, Narasimhan 2012б and others). However, there remains a number of methodological complications that have not been yet resolved. They are primarily associated with the fact that lexical analysis requires an ample and representative body of lexical data which, in most instances, cannot be retrieved from lexicographic sources alone. Therefore, researchers are faced with the necessity to compile task-specific questionnaires and to invite experts in multiple languages to interview the informants and to analyze the obtained materials. The high cost of such projects inhibits the possibility of detailed analyses of extensive semantic fields across a larger number of languages. As a consequence, researchers are compelled to either dramatically curtail the number of languages in the sample, or to reduce the granularity of the analysis. Meanwhile, the results of thorough and large-scale comparative analysis of lexical meanings could have high practical utility, along with their indisputable theoretical value: they could be utilised in manual and machine translation, and in improving the efficiency of second language teaching.

The present work is based on the premises of the frame-based approach to lexical typology developed by the Moscow Lexical Typological Group, MLexT (Rakhilina, Reznikova 2016), which, in its turn, stems from the traditions of the Moscow semantic school, see Апресян 1974. The key notion of this research paradigm is the concept of frame; it stands to indicate a minimal situation which is denoted by a specialised lexeme in a given language. Typological description of a semantic field consists in identifying the frames that constitute it (i.e. the types of situations that are covered by the lexemes of the field), and in describing the models of their lexicalization (i.e. how the situations converge within lexemes).

The set of frames is established by analyzing the combinability of words as attested in dictionaries and corpora, and is further elaborated and finalised via interviews with speakers of the language. The convergence of frames under individual lexemes is visualised by means of semantic maps which are similar to the maps used in grammatical typology (see Haspelmath 2003).

The present thesis proposes new methods for a u t o m a t e d collection and analysis of lexical typological data; **it is motivated by the methodology and theory** of the frame-based approach. The suggested methods aim to:

(1) promote faster and easier collection of initial data;
(2) uncover new patterns of convergence and divergence of meanings within lexemes, which could not be captured by the previously used methods;
(3) minimise the dependence of lexical typological analysis on the researcher's initial theoretical premises and his / her native language, as well as on the languages from which the data for the questionnaire is drawn.

The primary goal of the thesis is accomplished through the following series of **tasks**:
(1) formalise the procedure of frame-based lexical typological research, i.e. subdivide it into stages and formulate the problems to be addressed in each stage;
(2) suggest automated methods for implementation of each stage and test them;
(3) analyze the performance of the suggested methods and outline the directions for further development of the most efficient algorithms.

The algorithms for automated collection and analysis of lexical data proposed in the present thesis are based on the following **methods**: distributional semantic models (Baroni et al. 2013), cluster analysis (Everitt 2011), and formal concept analysis (Ganter, Wille 1999).

The **relevance** of the thesis is determined by the proven demand for automated lexical analysis that exists in both theoretical and applied linguistics.

The **scientific novelty** of the present research is warranted by the fact that lexical typology remains a largely understudied area, and by the relatively small number of projects developing computational methods of lexical analysis. The methods used in this thesis have not been applied to lexical typological problems before; we suggest exclusive algorithms for their implementation in lexical typological analysis.

The following propositions **are submitted for the defense**:
(1) The frame structure of a field is represented by overlapping clusters with distinct centres (or "foci');
(2) A draft version of a lexical typological questionnaire can be compiled from a monolingual corpus of texts by means of distributional semantic modeling and by the subsequent cluster analysis of the obtained distributional space;
(3) The questionnaire can be translated automatically with the aid of parallel corpora, machine-readable dictionaries, and online machine translation engines; it can be filled (at least partially) with data from parallel and monolingual corpora;
(4) Formal concept lattices can be used as semantic maps to visualise the relative distances between the frames of direct meanings of different lexemes as well as the systemic relations between the direct and the figurative meanings within a semantic domain.

The **theoretical significance** of the present work is defined by its contribution to the development of lexical typology in general and to the frame-based approach in particular. The results of the research provide new insights into organization of the semantic space of lexical meanings and propose new hypotheses about the degree of their comparability.

The **practical significance** of the thesis consists in the development of algorithms to optimise the process of lexical typological research and thus facilitate preparation of data for the

purposes of conventional and computational lexicography, second language teaching, and manual and machine translation.

All the experiments presented in this thesis are based on the **material** of the earlier studies of several semantic fields of qualitative features and verbs that were manually conducted by members of the MLexT group, such as the fields of 'sharp' (Кюсева 2012), 'smooth' (Кашкин 2013), 'straight' (Лучина 2014), 'thick' and 'thin' (Козлов et al. 2016), 'oscillation' (Шапиро 2015), 'falling' (Reznikova, Vyrenkova 2015), and several others.

Public demonstrations of the results. The major results of the research were presented at the following academic events: the 9th Conference on Typology and Grammar for Young Scholars (Saint-Petersburg, Russia, 2014), the Workshop on Lexical Typology at the University of Helsinki (Helsinki, Finland, 2014), the 16th April International Academic Conference on Economic and Social Development at the National Research University Higher School of Economics / NRU HSE (Moscow, Russia, 2015), the academic seminar with the visiting scholar Asifa Majid at the NRU HSE (Moscow, 2015), the International Conference on Computational Linguistics "Dialogue 2015" (Moscow, 2015), the conference "Problems of Computational Linguistics" (Voronezh, Russia, 2015), the 1st International Conference on Foreign Languages in Research and Education: Issues and Perspectives (Moscow, 2015), the international academic seminar "Computational Linguistics and Language Science" (Moscow, 2016), the poster session of the Typological School held by the School of Linguistics at the NRU HSE (Moscow, 2016), the 10th International Conference on Language Resources and Evaluation / LREC'16 (Portorož, Slovenia, 2016), the International Seminar on Perceptual Metaphor (Nijmegen, the Netherlands, 2016), the International Seminar on Motion Verbs (Paris, France, 2017), and the International Cognitive Linguistics Conference (Tartu, Estonia, 2017).

The contents of the thesis. The thesis consists of the Introduction, five Chapters, the Conclusion, the List of References containing 158 titles, and 5 Appendices. The Introduction provides a general overview of the thesis. The first section of Chapter 1 reviews the existing methods of comparative lexical analysis, including the state-of-the-art computational methods. The second section of Chapter 1 discusses the frame-based approach to lexical typology that forms the basis of this research. Chapter 2 presents the results of a series of experiments assessing the validity of the concept of frame that stands at the core of the frame-based paradigm. Chapters 3 – 5 introduce the methods that can be used to automate each stage of the analysis: designing the questionnaire (Chapter 3), filling the questionnaire with data from multiple languages (Chapter 4), and generating the semantic map (Chapter 5). The Conclusion sums up the major findings of the thesis and suggests the directions for further research.

SUMMARY OF THE THESIS

Chapter 1 "Introduction" contains a brief overview of the existing methods of lexical typological analysis. The fundamental premise of lexical typology states that lexical units in languages form a system, and that lexical systems can be compared across languages. The existing approaches in this area of linguistics primarily differ in their selection of parameters for comparison of lexical units. These parameters further guide the choice of data sources and how the data is processed and analyzed.

Five major lines of research can be defined in contemporary lexical typology:

(1) The experimental approach goes back to Berlin, Kay 1969 and their research on the typology of colour terms. In this methodology, questionnaires consist of extralinguistic stimuli, i.e. the comparison is carried out on the basis of perceptual characteristics of objects or situations.

(2) The theory of semantic primitives developed by A. Wierzbicka and C. Goddard (Wierzbicka 1985) stipulates that the meaning of any word in a natural human language is composed of a very limited set of universal semantic primitives (such as "I", "you", "something", "big", "think", etc.). Different combinations of universal primitives differentiate the meanings of individual words from each other.

(3) Dictionary-based approaches examine the submeanings listed in dictionary entries and attempt to identify patterns of their "colexification" (in the terminology of François 2008), i.e. their convergence within lexemes.

(4) The approaches that rely on parallel corpora (e.g. see. Viberg 2006, Wälchli, Cysouw 2012) use the contexts that are found across multiple languages as a proxy for a lexical typological questionnaire; words are differentiated by the contexts in which they can or cannot occur.

(5) The frame-based approach to lexical typology posits the existence of a universal set of the minimal lexical meanings (frames). It is considered that each semantic field features a peculiar set of frames, while different words cover various combinations of frames.

The first section of Chapter 1 looks at the first four of the above-mentioned approaches, discusses their weaknesses and strengths, and points out the general trend towards computerization of data collection and analysis.

The second section of Chapter 1 provides an in-depth discussion of the fifth approach that forms the groundwork of this research. The approach emerged from the traditions of the Moscow semantic school (see Апресян 1974); it suggests that comparative analysis of intralinguistic and crosslinguistic quasi-synonyms should be carried out in terms of their combinatorial properties. To compare words from different languages, the researcher should split their semantics into non-overlapping conceptual fragments, i.e. the types of situations in which these words are used. The types of situations correspond to different groups of contexts.

For instance, the semantics of the Russian adjective *tonkij* ('thin') can be represented as the following set of conceptual fragments:

- 'with a small diameter of the cross-sectional profile'; denotes the property of elongated objects ('pencil', 'rope', 'stick', etc.);
- 'with a small distance between the surfaces of an object' – about dimensions of flat objects ("layers"), such as a book, fabric or paper;
- 'low in intensity and high in pitch' – about the characteristics of sound; is used in the context of the nouns *zvuk* 'sound', *golos* 'voice', etc.

This set of elementary situations can also be used to identify the equivalents of the Russian adjective *tonkij* in translations. Cf., for example, the Chinese translations of this word:

- 'thin' + the name of an elongated object => *xì* (*xì gùnzi* – 'a thin stick');

- ‘thin’ + the name of a flat object => *báo* (*báo zhǐ* – ‘thin paper’), etc. (see Кюсева et al. 2013 for more detail).

Such situations are called **frames**. Frames are regarded as the minimal lexical meanings; i.e. individual lexemes cover various combinations of frames. Combinations of frames are not equally probable; some of the frames tend to converge under one lexeme while others have a stronger tendency to diverge into different lexemes. The patterns of frame convergence within a lexeme are visualised in lexical semantic maps.

This method of lexical semantic research has been applied to a wide range of lexical data, cf. Майсак, Рахилина 2007, Круглякова 2010, Кашкин 2013, Холкина 2014, Кюсева 2012. It proved to be efficient in establishing the frame structure of individual semantic fields and in comparing words across languages.

Research of lexical units in this paradigm consists of the following steps:

1. Compile the questionnaire (i.e. define the tentative set of frames) by analyzing the combinability of the lexemes of the given lexical field in 3-5 languages.
2. Finalise the set of frames: collect data from the other languages of the sample.
3. Describe and visualise the system in each language: draw the semantic map.
4. Analyze the types of systems observed in different languages.

To define the set of relevant frames, it is necessary to conduct a detailed analysis of data from dictionaries and corpora, and then extend it with data from interviews with speakers of the language. As the key objective of the research is to learn the combinability rules for each lexeme of the field, the questionnaires for the interviews are comprised of the contexts where the lexemes occur. This means that the finalised questionnaire has to be translated into each language of the sample.

Today, practically each of these steps is executed manually; this process is rather slow and requires meticulous and concerted effort from specialists in every language of the sample. The high cost and the need to recruit an expert for each subsequent language hinder large-scale investigations of representative language samples. As a consequence, the insufficient sizes of the investigated language samples cast doubt on the validity of the reported findings, particularly, on the adequacy of linguistic grounds for distinguishing frames as specific semantic units that claim the status of the minimal lexical meanings.

Chapter 2 “Verifying the concept of frame with distributional semantic models” describes a series of experiments intended to provide additional evidence in support of frames. The frame structure of a field is defined in terms of semantic similarity: the situations that belong to one frame display the closest semantic proximity, while the distances between situations from different frames can vary. The distances between the frames are represented in the semantic map of the field.

Semantic distances between frames are measured on the basis of typological data. Normally, the frame-based approach considers only relative distances, i.e. if lexeme L1 can cover frames F1 and F2, and lexeme L2 can cover frames F2 and F3, and there is no lexeme that could cover F1 and F3 without covering F2, it is argued that frames F1 and F3 are farther apart than F1 and F2 or F2 and F3. This configuration of frames is illustrated by the linear semantic map: F1 – F2 – F3.

We extended the pilot study by Кюсева 2014 and developed a formula for a more precise quantitative measurement of typological similarity between frames; the formula is based on the frequency of colexification of the minimal meanings. Each frame is expressed as vector *w* whose dimensions are represented by the lexemes that belong to the field in all the languages of the sample. If lexeme l_i can denote the frame, the value of the corresponding dimension is 1, otherwise 0. The typological distance between frames is computed as cosine similarity between the respective vectors (cf. a similar metric of similarity discussed in the recent study by Youn et al. 2016).

However, there exist other measures of semantic distances between lexical meanings. One of them (known as distributional semantic models, see Baroni et al. 2013) represents the meaning of a lexical unit (a word or a phrase) as a vector of its co-occurrences. Such semantic representations are used for a wide range of purposes, some of which are quite similar to our task (e.g. in word sense disambiguation and in matching contexts with the most appropriate of a number of quasi-synonyms). To the best of our knowledge, distributional semantic methods so far have not been implemented in typology. However, it is quite reasonable to assume that if it is true that frames are equivalent to the minimal lexical meanings, then the distributional distances between them can be expected to correspond to the typological distances.

We conducted several experiments with the fields of qualitative features 'sharp' and 'smooth'; the sets of frames and the data for computation of typological distances were drawn from the Typologically Oriented Database of Qualitative Features (see Кюсева et al. 2013a).

For each frame we selected several illustrations (or 'micro-frames'; e.g. the frame 'instrument with a sharp functional end-point' was illustrated by 'sharp needle', 'sharp arrow' and 'sharp spear' from the field of 'sharp'. Each micro-frame was aligned with its two-word Russian counterpart (*ostraja igla*, *ostraja strela*, *ostroe kopje*); for each of these word combinations the co-occurrences vector was obtained.

The distributional models in our experiments were developed with the parameters described below.

1. **Dimensions:** 10,000 most frequent content words lemmas (in the main subcorpus of the Russian National Corpus (RNC))
2. **Values of the dimensions:** the frequency of content words in the window of ±5 around the lexical unit for which the vector is retrieved
3. **Distances between the vectors** are computed as cosine similarity measure
4. **Training corpus:** different combinations of the main subcorpus of the RNC (approx. 220m tokens), the newspaper subcorpus of the RNC (approx. 200m tokens), and the ruWaC corpus of texts collected from the Internet (around 1bn tokens)
5. **Vector processing:** vector weighting (none vs. PPMI vs. PLMI vs. PLOG vs. EPMI) and dimensionality reduction (none vs. SVD to 300 dimensions)
6. **Type of vector representation of phrases**: the observed vector (when the focal word combination is regarded as a single lexical unit) vs. the composed vector (when the vector of the phrase is computed from the vectors of its constituent words according to one of the following models of composition: additive, weighted additive, multiplicative, dilation, lexical function, practical lexical function, and PLF)

Thus, two vector representations were obtained for each micro-frame: the typological and the distributional vectors. After that, we calculated the typological and the distributional distances for all the possible pairs of micro-frames within each field and computed the Pearson correlation coefficient between the two metrics.

Both of the fields under study demonstrated very high correlation coefficients (0.766 for the field 'sharp' and 0.946 for the field 'smooth'). As some of the parameters of the distributional models were run with various settings, it is important to note that the best results for both fields were achieved with the same set of settings: the main subcorpus of the RNC as the training corpus, vector weighting with PPMI, dimensionality reduction to 300, and the PLF (practical lexical function, see Paperno et al. 2014) model to vectorize phrases as a composition of the vectors of their constituent words. It also should be noted that the best results were shown for the direct meanings of the adjectives; adding metaphoric frames significantly decreased the overall performance: the Pearson correlation coefficient in this case was 0.462 for the field 'sharp' and 0.604 for the field 'smooth'. A possible explanation is that direct meanings possess a more distinct and, importantly, a more predictable frame structure than figurative meanings; although figurative meanings are motivated by direct ones, the distribution of figurative meanings in the language is less even.

Thus, despite the commonly held opinion (e.g. see Bullinaria and Levy, 2012), the quality of the model does not increase in proportion to the size of the training corpus; in our case, the small yet well-balanced main subcorpus of the RNC yields a higher result than the joint corpus of approximately 1.44bn tokens comprised of the main and the newspaper subcorpora of the RNC, and the ruWaC corpus (cf. a similar observation in Kutuzov, Kuzmenko 2015).

The main subcorpus of the RNC proved to be sufficient enough for obtaining high-quality vector representations of individual lemmas. At the same time, even the joint training corpus did not appear to be large enough to produce reliable combinability profiles of two-word phrases; the quality significantly increases (as against the co-occurrence vectors) when we apply any of the composition models.

Evidence from only two semantic fields does not allow us to make definitive conclusions with a high degree of confidence; therefore, we performed two further experiments. In the first experiment we implemented the best parameters of the distributional models from the two previous experiments with adjectives; but this time we applied them to a verbal semantic field ('oscillation')[1]. In the second experiment we trained the models on a new corpus, the English ukWaC, and tested them on the field of 'sharp'. Thus, the invariable typological distances were compared against the distributional distances between the respective English word combinations, e.g. *sharp needle*, *sharp spear*, *sharp arrow*, etc. Both of the experiments yielded high Pearson correlation coefficients: 0.7 for 'oscillation' and 0.668 for 'sharp'. These results offer further support for the hypothesis that the frame structure of a field can be roughly outlined using data from only one language, and it is irrelevant which language provides the data.

Finally, the correspondences between the typological and the distributional spaces are convincingly demonstrated by their visualizations. We used multidimensional scaling to project

[1] We thank Maria Shapiro who provided her typological data for this experiment.

the two spaces of each field onto planes, where each frame is coded with its specific colour. Figs. 1-3 show the plots for the typological and the distributional spaces of the field ‘sharp’. Green markers in all the plots depict the frame ‘instrument with a sharp functional edge’; the frame ‘instrument with a sharp functional end-point’ is coded with blue; yellow represents the frame ‘elongated object’, and red stands for the frame ‘object with prickly surface’. It should be emphasised that these clusters were not based on the obtained maps; they were initially determined on the basis of the typological studies carried out by the MLexT group. From here on, we will refer to the ‘green’, ‘blue’, ‘yellow’, and ‘red’ clusters as the respective four frames of the field ‘sharp’.

Figs. 1-3 demonstrate a remarkable effect. Visualization of the typological space (Fig. 1) accurately reflects the frame structure of the field[2]. Visualization of the distributional space, on the contrary, captures only the juxtapositions that are lexicalised in the given language. For example, the map in Fig. 2 was generated from the Russian data, and the contexts for the adjective *koljučij* (‘prickly’) stand out from the rest, while the frames of the adjective *ostryj* (‘sharp’) form a smooth continuum. Fig. 3 depicts visualization of the distributional space for the field ‘sharp’ built on the basis of the French corpus (for visual clarity, it contains only the frames that are not differentiated in Russian). In the French data, the frame for ‘instrument with a sharp functional edge’ is distinctly juxtaposed to the frames ‘instrument with a sharp functional end-point’ and ‘elongated object’ because the first of these frames is denoted by the adjective *tranchant* while the other two correspond to *pointu*, and this juxtaposition is lexicalised in French.

[2] Note that multidimensional scaling was successfully used in typology for automated generation of semantic maps (see Croft and Poole 2008, Wälchli and Cysouw 2012, and others). A more detailed discussion of multidimensional scaling will follow in Chapter 5.

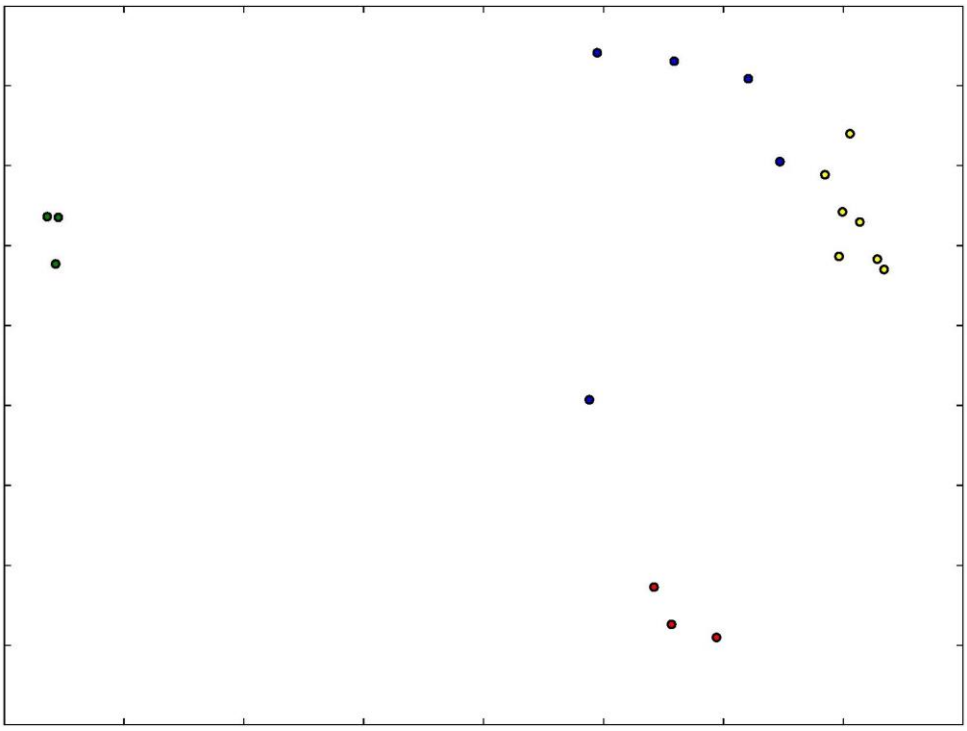

Fig. 1. Visualization of the typological space of the field ‘sharp’

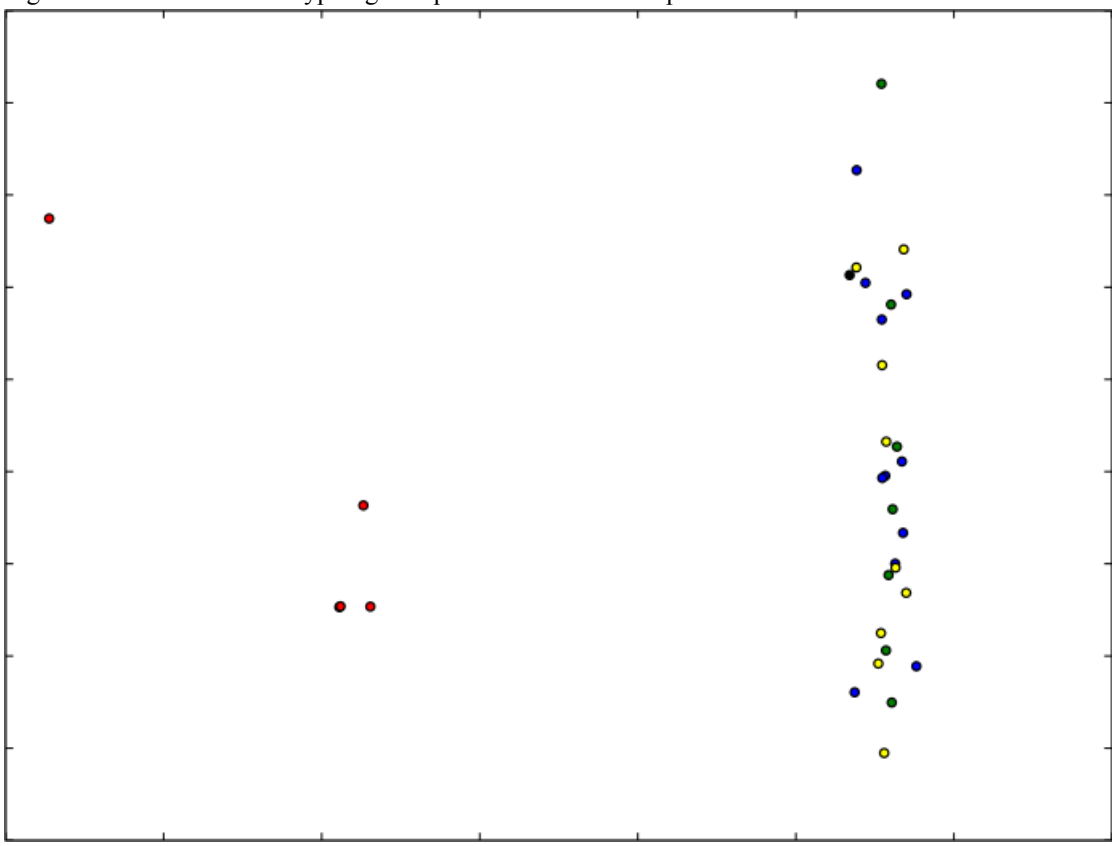

Fig. 2. Visualization of the distributional space of the field ‘sharp’ generated on the basis of the Russian corpus

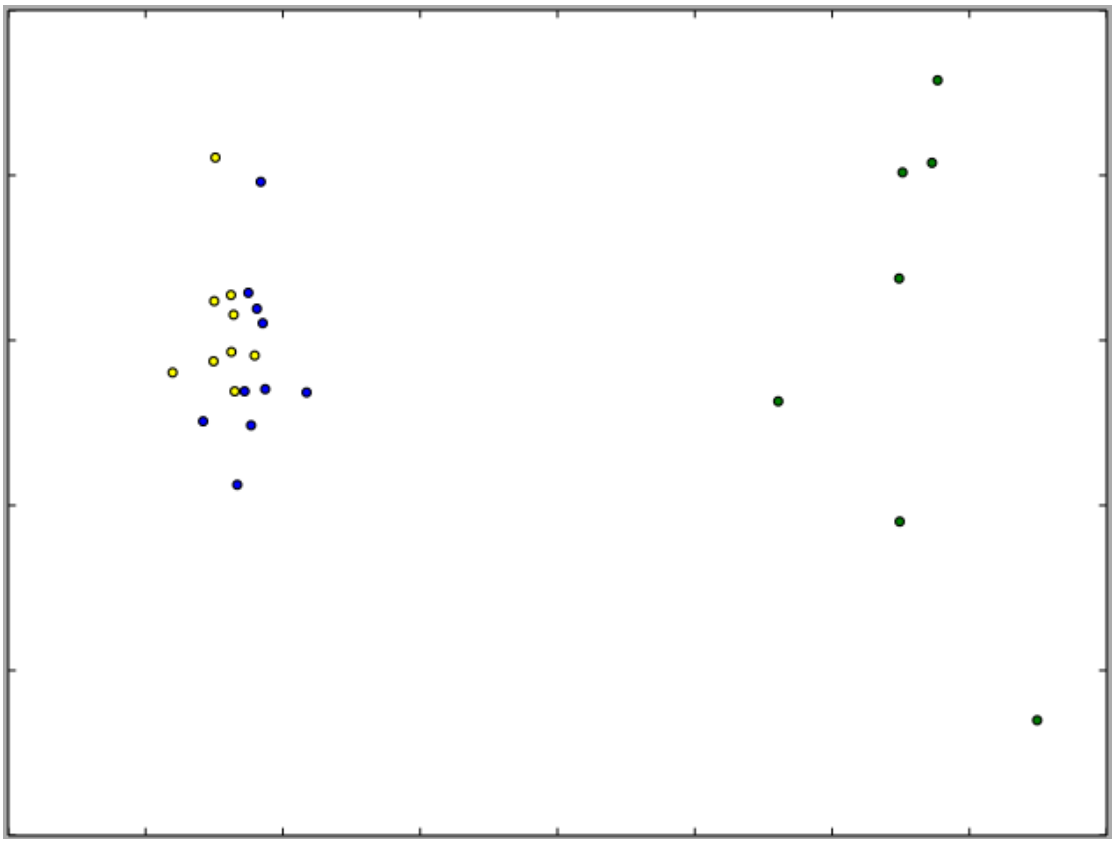

Fig. 3. Visualization of the distributional space of the field ‘sharp’ generated on the basis of the French corpus (without the frame ‘object with surface that pricks’)

However, the picture changes dramatically if we do not indiscriminately plot all the objects in a vector space; instead, we determine the nucleus of each frame, and plot only these nuclei. We computed the means of each dimension in every “cluster” to define the centre of each frame, and mapped these new reduced spaces onto a plane. As demonstrated in Fig. 5, the maps produced with this method from the data of only one language are identical to the conventional discrete semantic maps (Fig. 4).

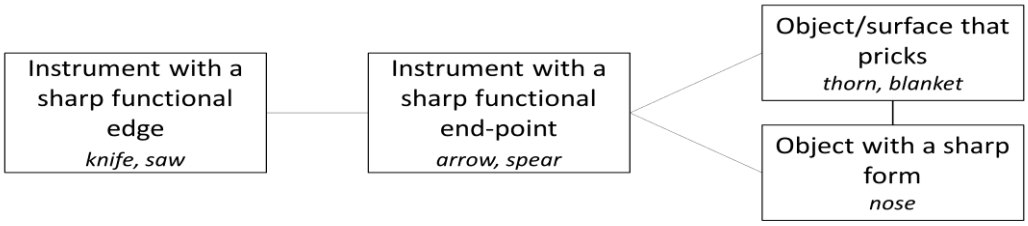

Fig. 4. Semantic map of the field ‘sharp’ compiled manually from typological data

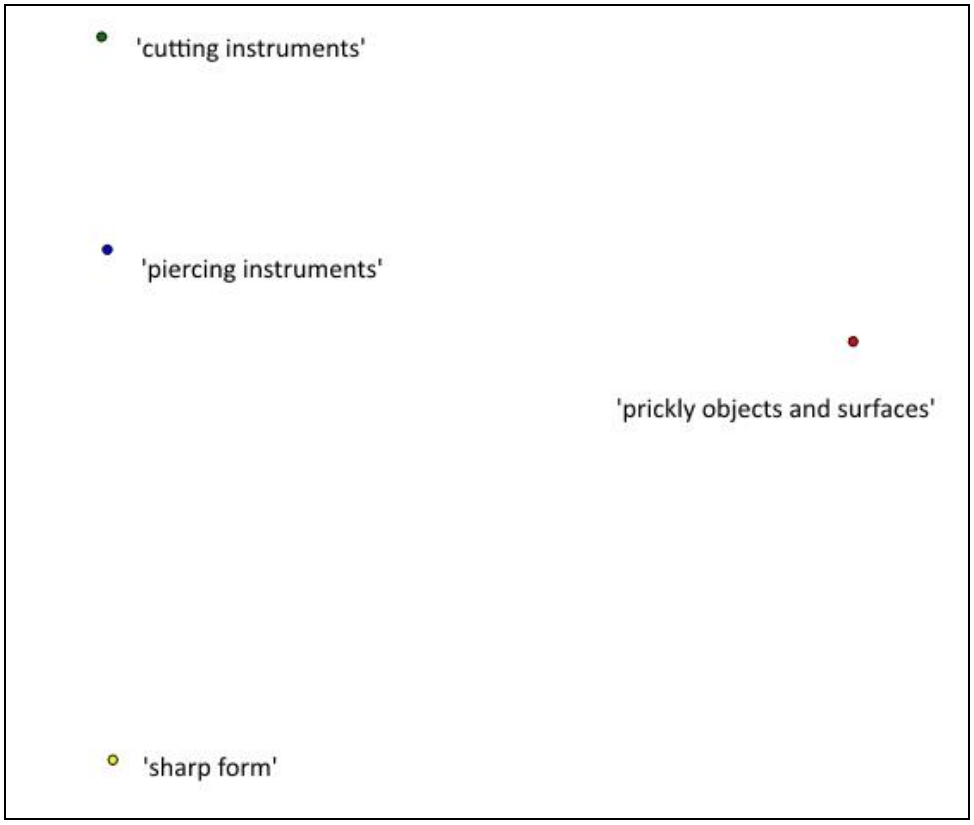

Fig. 5. Automatically generated semantic map of the field ‘sharp’: the mapping of the vector space constituted by the central objects of each frame cluster

There are three principal conclusions that follow from our findings:

(1) Significant correlation is observed in all the four experiments between the typological and the distributional spaces; therefore, accurate manually collected typological data can be used to evaluate the quality of distributional models. Such metric has several advantages over the other existing methods (e.g. comparing distributional distances against spontaneous judgements elicited from speakers of the language, or against the length of the path in the tree of a thesaurus); the key advantage of our metric is its objectivity. Its major drawback is primarily associated with the lack of reliable typological data; however, we expect that advancements in algorithms for automated data collection will help to resolve this problem in the short term.

(2) Our results offer further evidence that corroborates the presence of a linguistically grounded semantic reality behind the concept of frame. Nevertheless, it would be an oversimplification to regard frames as points in the semantic space; this approach was motivated by manual processing of the data. In fact, the frame structure of a semantic field seems to be continuous rather than discrete; however, this continuum of meanings does have distinct focal points (cf. Кибрик 2012); these are the frames that in most cases define the principles of lexicalization of the field.

(3) Distributional semantic methods and data from one language are sufficient to produce a rough outline of a semantic field; moreover, the experiment with the English corpus suggests that the choice of the initial language does not affect the final result.

We have provided additional evidence that supports the concept of frame and demonstrates that it is well-grounded both in theory and practice; now we move on to the discussion of the possible methods that can be used to automate the stages of a research project in frame-based lexical typology.

Observation (3) from the previous chapter allowed us to develop a method for generating lexical typological questionnaires from the data of one language. The method is described in Chapter 3 "Automated development of questionnaires with distributional semantic models".

The suggested algorithm produces draft questionnaires for typological studies of adjectives and other one-place predicates, such as the verbs of motion, sound, or state.

The algorithm consists of the following steps:

1. Compile the list of nouns that co-occur with the adjectives / verbs under examination (in the main subcorpus of the RNC);

2. Obtain the vector of co-occurrences for each word combination;

3. Cluster the resulting vector space by means of hierarchical clustering.

4. Select the three central objects in each cluster and remove the clusters that contain less than three objects.

Steps 2-4 are always applied irrespective of the part of speech. Step 1, however, depends on the format of the minimal diagnostic test required for the reference word. We presumed that in order to establish the meaning of adjectives and other one-place predicates it is sufficient to look at the noun that occupies the position of the word's only actant. Therefore, in the case of adjectives we retrieved the nouns that occur to the right of the reference word in the main subcorpus of the RNC; for verbs we retrieved the nouns in the Nominative Case that occur either to the left or to the right of the reference word.

The algorithm was developed and tested on four fields of qualitative features ('sharp', 'smooth', 'straight', and 'thick') and one verbal field ('oscillation'). For each of the fields we evaluated the recall and the precision of the resulting questionnaire. The recall was measured as the percentage of frames that are represented in the questionnaire by at least one illustration; the precision was estimated as the purity of the clusters.

	Recall	**Precision**	**F-measure**
'sharp'	0,733	0,827	0,777
'straight'	1	0,817	0,899
'smooth'	0,8	0,675	0,732
'thick'	1	0,884	0,938
'oscillation'	0.882	0.762	0.818

Table 1. Quantitative evaluation of the algorithm's performance

Table 1 displays the measures of the algorithm's performance in each of the tested fields. As can be seen, the overall performance is fairly good, but some of the fields yield much better results than the others. This discrepancy can be accounted for by a number of factors.

The first factor is the frequency of the reference lexeme; higher frequency (i.e. a greater number of occurrences of the lemma in the corpus) results in better performance. The low F-measure for the field of oscillation can be attributed to the low frequency of its lexemes and the subsequent inadequacy of the vector representations and clustering.

Secondly, the quality of performance depends on the number of frames in the field. Fields are better clustered into semantically homogeneous groups when they contain fewer frames. This explains the high quality of clustering of the field 'straight': it contains seven frames, and each frame is represented by a large number of contexts.

Thirdly, the quality of a questionnaire is affected by the nature of the oppositions that organise the semantic structure of the field. The suggested method for automated questionnaire generation groups contexts according to their taxonomic classes. For example, the Russian words *potomok* ('descendant') and *predšestvennik* ('predecessor') from the class of human beings are placed into one cluster of the field 'straight', while *alleja* ('parkway') and *dorožka* ('pathway') belonging to the class of extended areas are referred to another cluster. In most cases this leads to the desired partitioning of contexts into frames. However, it is not always the case that frames in a field are juxtaposed to each other in accordance with the taxonomic classification of nouns; in some cases it is the topology of the object that matters. For example, the two frames of the field 'sharp' – 'instrument with a sharp functional edge' (e.g. a knife or a sword) and 'instrument with a sharp functional end-point' (e.g. a needle or a bradawl) – belong to the same taxonomic class of nouns (instruments), but differ in the topological characteristics of the objects they describe: the first one is characterised by a linear segment while the second one – by a point-like segment. Differences of this kind are captured by the algorithm much less reliably.

The factors enumerated above are not equally relevant. For example, the algorithm delivered the best performance for the field 'thick' despite the fact that this field is essentially structured according to the topological classification of objects. This effect may be due to the high frequency of the adjectives of this field and the low number of frames in its semantic structure. Besides, the topological and the taxonomic classifications of nouns often correlate with each other, thus contributing to the purity of clustering. For instance, many body parts belong to the topological class of elongated objects (e.g. thick fingers, arms, or legs), while pieces of clothing often belong to the topological class of flexible layers (e.g. a thick jacket, coat, or sweater).

Chapter 4 "Methods for automated data collection" describes the approaches that facilitate collection of data; two tasks are addressed at this stage: (1) translating the minimal contexts from the questionnaire; and (2) filling in the questionnaire with data from the relevant languages. We experimented with the fields of qualitative features ('sharp', 'smooth', 'thick', and 'thin'); therefore, task (1) consisted in translating the list of adjectives belonging to the field along with the list of nouns that may potentially co-occur with them.

Translation of adjectives is no trivial matter. Traditionally, the task of translation (including machine translation) is regarded as either matching a context with the most suitable lexeme, or finding the most frequent / accurate translation equivalent of a lexeme, or generating the best equivalents for each of the meanings of the original word. Our task is different from those above; we need to obtain the adjectives that translate the original words only in the contexts that correspond to their direct meanings. For example, among the English translation equivalents of the Russian word *ostryj* we would like to see the adjectives *sharp* and *pointed* but not *critical* or *urgent* (cf. *ostraja nexvatka* ('critical shortage'), or *ostryj vopros* ('urgent matter')).

We tested several algorithms (which are described in detail in the main text of the thesis) and opted for the method based on the machine-readable dictionaries of the FreeDict group. The advantage of these dictionaries is that translations are aligned with word meanings; our algorithm picks the translation equivalents of only the first meanings and then double-checks them by means of back-translation, when the candidate is translated back into the source language. The candidate is added to the final list only if the adjective that corresponds to its first meaning is contained in the initial list. Nouns are translated in a similar manner, with a slight modification: if a noun is not present in the FreeDict dictionary, it is translated with a machine-readable dictionary by Yandex company.

After that the questionnaire is converted into the tabular format where columns are headed as adjectives and rows are headed as nouns. The table is filled with data from the available corpora: if an adjective co-occurs with a noun in a corpus, we compute the mutual information for this pair. Combinations with the negative value of mutual information are considered to be random and are excluded from the final questionnaire.

Chapter 5 "Automated generation of semantic maps with formal concept lattices" describes the methods used to automate the final stage of the analysis. Special focus is placed on the theory of formal concept analysis (Ganter, Wille 1999) which introduces a special kind of diagrams known as formal concept lattices (FCLs). We maintain that such diagrams can be used in linguistic research as a new type of semantic maps.

FCLs are based on the so-called formal contexts. Formal context $K = (G, M, I)$ is a set of objects (G), a set of attributes (M), and the binary relation (I) between the objects and their attributes. A formal concept is a pair (A, B) where A is a subset of G and B is a subset of M so that B contains all the attributes that characterise the objects in A, and A contains all the objects that feature the attributes from B within a given formal context. FCLs represent data as a hierarchy of formal concepts where concepts are ordered from more generic to less generic (those that cover a smaller number of objects).

In our case, the objects are represented by lexemes, and attributes are represented by frames. A lexeme and a frame form an incident pair if the lexeme covers the frame. We experimented with ten fields of qualitative features ('sharp', 'soft', 'smooth', 'rough', 'hard', 'empty', 'thick', 'thin', 'high' and 'low') and with the verbal field of falling.

To the best of our knowledge, this method has not been used in linguistics before (Priss 2005 is one of the few exceptions); our experiments demarcated the limits of its applicability to lexical typological research. We established that FCLs can be used as is, without any modifications, to represent the fields with the linear structure – such that form a chain-like configuration in the conventional semantic maps (Frame 1 - Frame 2 - Frame 3). In this case, the arrangement of the nodes in a lattice corresponds to the arrangement of frames in a conventional semantic map; however, lattices are generated automatically while conventional semantic maps are designed manually.

FCLs go beyond automating the design of maps: they open up new opportunities for linguistic research. Firstly, the hierarchical arrangement of nodes in a lattice makes it possible to show in one diagram all of the lexicalization strategies that are available for a given field (in contrast to the traditional semantic mapping technique where convergences of frames are depicted separately for each language), see Fig. 6. This, in turn, considerably simplifies the

typological analysis: it appears that some of the combinations that are considered admissible in a conventional semantic map are never or very rarely realised, while others are highly frequent. For instance, the FCL for the field ‘sharp’ indicates the two most frequent strategies in our sample: the dominant, in which all the major frames are covered by one lexeme, and the binary, in which one lexeme denotes instruments with a sharp functional edge (knives, saws, razors, etc.) and the other lexeme covers instruments with a sharp functional end-point (spears, arrows, etc.) and elongated objects (a nose, the toes of a boot, etc.). Interestingly, the conventional semantic map does not capture this tendency; it simply places the frame ‘sharp (used about instruments with a sharp end-point)’ between the other two frames, which implies that it is equally likely to converge with the frame ‘sharp (about elongated objects)’ and ‘sharp (about instruments with a sharp edge)’ (see. Fig. 4 above).

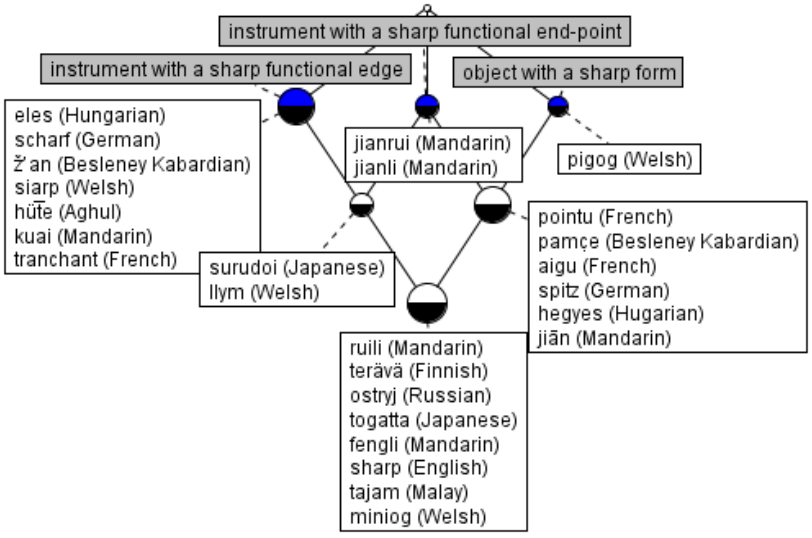

Fig. 6. FCL for the semantic field ‘sharp’

Secondly, due to the hierarchical organization of nodes, both direct and figurative meanings can be visualised in the map. At the same time, the map preserves the schematic representation of the degree of proximity between the direct meanings, because all the corresponding nodes are located at one level of the lattice, while the nodes of the figurative meanings occupy the level below. This configuration manifests the relations between the original and the resulting semantics of the lexemes that underwent metaphorization, and visualises the models of the relevant semantic shifts (see Fig. 7); this fully conforms to the observations on the nature of the juxtaposition between direct and figurative meanings proposed by Рахилина, Резникова 2013.

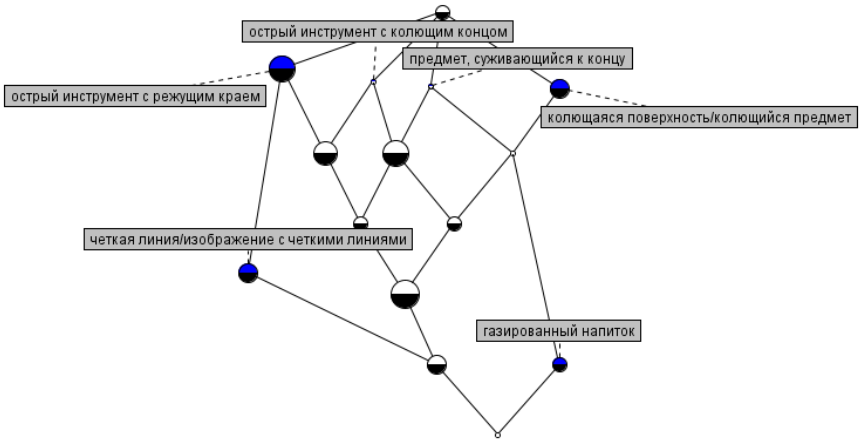

Fig. 7. FCL depicting some of the relations between the direct and the figurative meanings in the semantic field ‘sharp’

When the topology of a semantic field is more complex than linear, the graphic representation of the FCL becomes illegible because FCLs are more informative than conventional semantic maps.

To sum up, the suggested methodology:

(1) can be used to automate generation of (advanced) semantic maps for linearly organised fields or for linear segments of fields with a more complex structure;

(2) can serve as the basis for a subsequent algorithm for automated generation of semantic maps that would convert FCLs into conventional graphs. However, to implement such an algorithm, it is necessary to decide which information should be kept, and which regularities that are captured by formal concept analysis may be discarded when transforming an FCL into a graph.

The Conclusion summarises the key outcomes of the research:

(1) It is the first time that quantitative methods of collection and analysis of data are implemented within the frame-based approach to lexical typology;

(2) We provided new, quantitative evidence that the lexical typological concept of frame is grounded in the linguistic reality;

(3) It was demonstrated that frames are overlapping clusters with distinctly pronounced centres, i.e. prototypical situations.

(4) We proposed and described a formalised procedure of a frame-based lexical typological research project; the project is divided into several stages each of which deals with specific problem;

(5) We developed algorithms that automate each stage of the project. The proposed algorithms perform the following functions:

   a) generate the draft of a lexical typological questionnaire;

   b) translate the draft into multiple languages and fill it with data from available corpora;

c) plot the collected data onto the semantic map; the map visualises relations between the frames of both direct and figurative meanings;

(6) We performed an in-depth analysis of the obtained results and suggested directions for further development of the methods that proved most efficient.

In summary, the present research demonstrates that implementation of the latest developments opens up new vistas for lexical typological analysis. The new methods:

(1) drastically accelerate the research process and dramatically broaden the scope of linguistic data available for investigation;

(2) offer the maximum possible degree of independence from the theoretical premises initially held by the researcher and from their native language;

(3) uncover new patterns in organization of lexical meanings.

The following **papers have been published** discussing the subject of the present thesis:

1. Constructing of Lexical Typological Questionnaire for Adjectives and Intransitive Verbs with Distributional Semantic Models // RSUH/RGGU Bulletin, Moscow: 2016. Issue 9 (18). P. 140-150. (in Russian)
2. *A Fantastic Conference, a Terrible Talk*: Derivation of Evaluative Meanings in Russian Adjectives // Moscow State University Bulletin. Series 9: Philology. 2016. Issue 6. P. 178-192. (in Russian)
3. Grammatičeskaja polisemija skvoz' prizmu leksiki: instrumentalis v besleneevskom dialekte kabardino-čerkesskogo jazyka [Grammatical Polysemy through the Lexical Lens: the Instrumental Case in Besleney Kabardian] // In: ACTA LINGUISTICA PETROPOLITANA. Transactions of the Institute for Linguistic Studies of the Russian Academy of Sciences / Ed. N. Kazansky. Vol. 12. Part 2. Proceedings of the 10th Conference on Typology and Grammar for Young Scholars (2013) / Ed. D. Gerasimov. Saint-Petersburg: Nauka Publishers, 2016. P. 665-678. Co-authored with P. Arkadiev and M. Kyuseva. (in Russian)
4. Verbs with the Semantics of Falling and Locative Meanings in Kuban Kabardian // Proceedings of Voronezh State University. Series: Linguistics and Intercultural Communication. 2016. Issue 2. P. 79-85. Co-authored with M. Kyuseva. (in Russian)
5. Development of the Lexical Typological Inventory Based on the Distributional Semantic Models // Proceedings of Voronezh State University. Series: Linguistics and Intercultural Communication. 2015. Issue 3. P. 127-132. (in Russian)
6. Formal concept lattices as semantic maps // Proceedings of the 1st International Workshop on Computational Linguistics and Language Science (CLLS 2016), CEUR-WS.org, Eds. D. Ilvovsky, E. Chernyak, A. Vybornova, D. Skorinkin. 2017. Co-authored with S. Obiedkov.
7. Verbs of Sounding as the Material for the Theory of Semantic Models // In: Verbs of Animal Sounds: the Typology of Metaphors / Eds. T. Reznikova, A. Vyrenkova, B. Orekhov, and D. Ryzhova. Languages of Russian Culture (LRC) Publishing House, 2015. P. 325-343. Co-authored with E. Rakhilina and M. Kyuseva. (in Russian)
8. Verbs of Animal Sounds in Hindi // In: Verbs of Animal Sounds: the Typology of Metaphors / Eds. T. Reznikova, A. Vyrenkova, B. Orekhov, and D. Ryzhova. Languages of Russian Culture (LRC) Publishing House, 2015. P. 141-155. Co-authored with E. Bessonova (Kozlova). (in Russian)
9. Verbs of Animal Sounds in Bzhedug Adyghe // In: Verbs of Animal Sounds: the Typology of Metaphors / Eds. T. Reznikova, A. Vyrenkova, B. Orekhov, and D. Ryzhova. Languages of Russian Culture (LRC) Publishing House, 2015. P. 233-244. Co-authored with M. Kyuseva. (in Russian)
10. Typology of Adjectives Benchmark for Compositional Distributional Models // Proceedings of the Language Resources and Evaluation Conference, 2016. P.1253-1257. Co-authored with M. Kyuseva and D. Paperno.