# Thesaurus of Russian adjectives, organized into an array of linearly ordered groups of semantically close words, densely packed on the flat surface

**Vvedensky Victor**
RNC Kurchatov institute/ Moscow,
Russia
victorlvo@yandex.ru

**Abstract**

We split thesaurus of Russian adjectives into an array of semantic categories including words with close meaning. For compilation of the semantic categories we used Russian-English translations as the measures of proximity of word sense. The words can be divided into three domains, nearly equal in size, containing adjectives describing positive, neutral and negative qualities. Words in the groups are arranged in ordered lists, and these groups can be packed densely side by side. The whole thesaurus can be mapped onto a nearly circular patch of surface divided into three sectors. We conjecture that the emerging structure reflects representation of words in the human cortex.

**Keywords:** Semantically close adjectives, word ordering, representation of words in the cortex.
**DOI:** 10.28995/2075-7182-2021-20-1233-1238

# Тезаурус русских прилагательных, представленный в виде множества групп упорядоченных семантически близких слов, плотно-упакованных на плоской поверхности

**Введенский В. Л.**
НИЦ Курчатовский институт
/Москва, Россия
victorlvo@yandex.ru

**Аннотация**

Тезаурус русских прилагательных представлен в виде плотноупакованного множества групп слов объединенных по близости смысла. В качестве меры семантической близости слов использовалось количество русско-английских переводов для каждого слова. Тезаурус распадается на три домена, примерно одинаковой величины, включающие прилагательные, описывающие положительные, нейтральные и отрицательные качества. Слова в группах сведены в упорядоченные списки, которые могут быть размещены рядом друг с другом. Весь набор прилагательных отображается на часть плоской поверхности в виде круглого пятна, состоящего из трех секторов. Мы полагаем, что образованная структура отражает представление слов в коре мозга человека.

**Ключевые слова:** Семантически близкие прилагательные, упорядочивание слов, представительства слов в коре мозга

## 1 Introduction

Researchers in computer science try to allocate words in a metric semantic space based on their meaning [1]. The array of words acquires geometric properties in this space [2]. Large lexical databases, like WordNet [3], group words into sets of cognitive synonyms expressing distinct concepts. The words are interlinked by means of conceptual-semantic and lexical relations. Most often these abstract spaces are visualized as clusters of semantically close words

similar to inflorescences formed around local basic word. The whole thesaurus looks like a large complex graph.
Neuroimaging studies face the same problem: how the words (or their representations) are allocated in the neural tissue in the human brain? Experiments show that the words are widely scattered across nearly the whole area of both cerebral hemispheres [4]. Cortical space is two-dimensional (though folded into numerous fissures) therefore one can use less elaborate techniques, then the computation of conceptual-semantic interrelations in the multidimensional space. We adopted dictionary approach to establish interrelations between Russian adjectives. Adjectives are a finite set of words belonging to a separate lexical category. Words are massively interrelated due to imprecise meaning of the most of them. The vague meaning of any word is reflected in the abundance of synonyms and usual ambiguity of the translation into foreign language. We used this ambiguity to organize all adjectives on the plane.

We addressed this problem during our experimental study of spoken word perception. It is very important to choose proper sets of words taken for particular measurement with human subjects, since individual uncontrolled variability of perception often makes results difficult to interpret. There are numerous methods to study spoken word perception [5], we measured the time needed to recognize heard words. The words presented to listener followed each other nearly at a pace of usual conversation. Our experiments are described in [6, 7]. We observed regularity in our data when the words presented for recognition in a single session were close in meaning. For each new experiment we compiled a group of adjectives with different semantic content. The words in the groups were invariably recognized by the listener with different delay and these words were ranked from the most quickly recognized to the "slowest" one. This inspired search for all possible groups of that type for experiments and we managed to cover nearly all thesaurus of Russian adjectives.

## 2 Results

We used translation services Google Translate, Reverso Context and Translate.academic.ru [8–10] to construct a table which contains translations between Russian and English adjectives. If necessary, we used Oxford Dictionaries [11] to clarify the uncertainty of the use of a particular translation. We accumulated 6081 Russian and 7364 English adjectives which form an interconnected array of words sharing common translations. Frequency dictionary for Russian [12] contains more adjectives though these extra words have only single translation or form small isolated groups. They fall out of interconnected array of adjectives.

Starting from any common Russian adjective we compiled a table including Russian words linked to the initial one through English translations. An example of such table is shown in Fig.1. During manual accumulation of words the table grows and the distribution of translations in the table (points) is initially highly scattered. After addition of several words computer performs ordering of the words in both Russian and English parallel lists. The procedure is the following. Each Russian word is specified by a set of numbers which shows positions of the corresponding translations in the English list. We take the average of these numbers as a single measure, which becomes running rank for this word. The same holds for each English word. After putting several words into a group the computer alternately orders Russian and English lists by the running rank of words. The result in graphic form is monitored visually. After addition of few new words into the group the procedure is repeated. Usually three or four iterations are sufficient to order the list with added words. The supervisor observes gradual "condensation" of points at the diagonal of the table. Scrutiny of the emerging ordered lists of adjectives by the supervisor at a certain stage reveals word, which clearly falls out of

line and has meaning evidently different from the majority of words in the group. This stems from the polysemy of words. A new group around this outlier word has to be organized.

At the early stage of generation of word lists with different meaning one can see that the emerging groups of adjectives can be separated into those which describe emotionally positive, negative or neutral qualities. Supervisor identifies the affiliation of each new group. We find nearly equal number of positive, neutral and negative semantic categories and of words in these categories. These groups differ in size ranging from about 60 to just a few words. Distributions of group size (number of semantically close words) are nearly identical in positive, neutral and negative domains. This implies some general background which controls the number of words the language "invents" for use in a specific area of human life. The words which are neighbors in the ordered group are close in their meaning, since they have common translation into another language. For all groups we observe that the word meaning gradually changes along the list, so that the words on the ends may have quite different sense. Nevertheless, any group of adjectives we compiled in this way can be considered as related to a certain distinct concept.

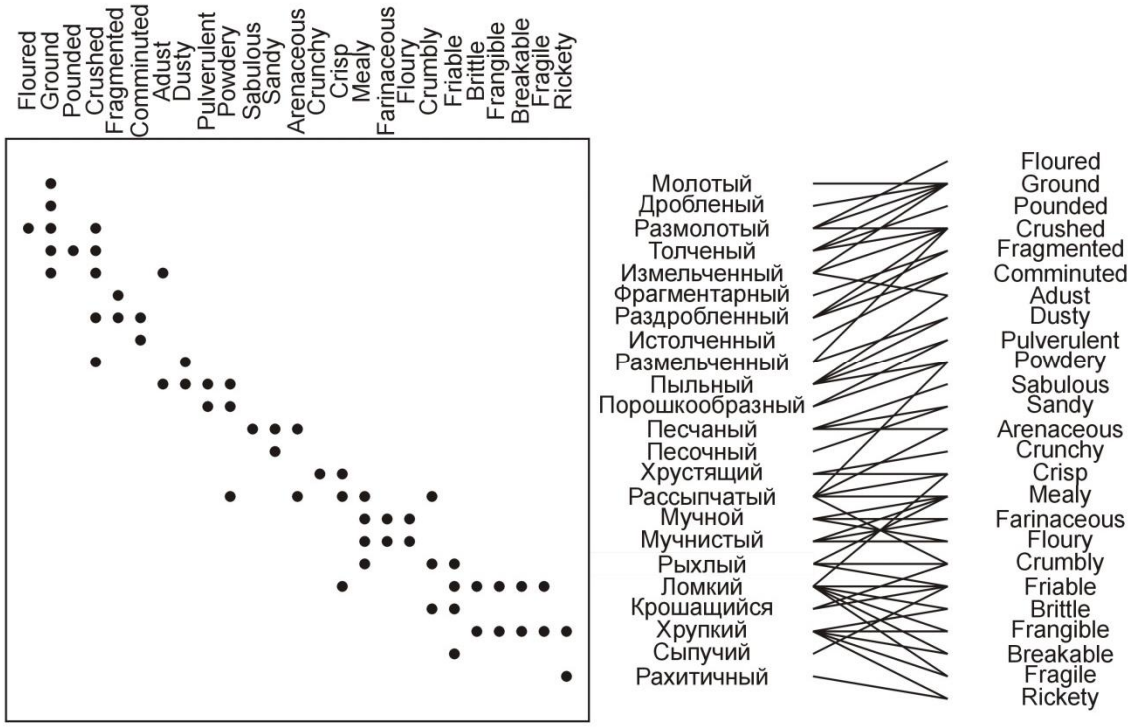


Fig.1. Russian-English translations in the group of semantically close words indicated as the points in the table. The points cluster tightly around the diagonal of the table and each word finds proper place in the Russian or English list. Right side of the plot shows parallel Russian and English ordered lists where allowed translations are indicated as lines stitching two lists together. The lines correspond to the points in the table.

Accumulating the translations we did found that the number of translation per word is quite robust characteristic of the whole thesaurus. Hundreds of Russian adjectives have up to six translations into English, while there are few words with up to 20 translations. That large number of translations provides reliable number of interconnections for words in the groups similar to the one shown in Fig.1. Translations "stitch together" parallel groups of words in two languages which can be attributed to a certain concept. Translations seem to be a reliable

basis stabilizing the emerging groups or semantic categories, since the number of translation per word follow quite robust mathematical dependence shown in Fig.2. The number of words falls exponentially with growing number of translations per word.

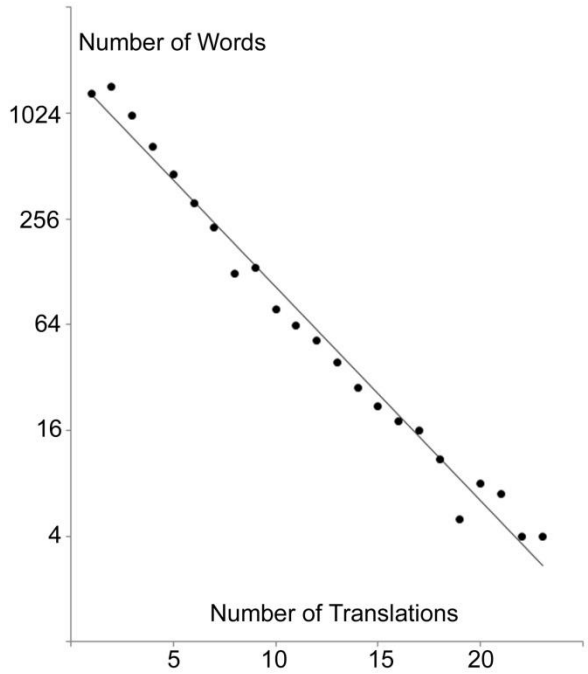


Fig.2. Number of Russian adjectives which have number of translations into English indicated on the horizontal axis. The scale of the vertical axis is logarithmic.

Each word finds proper semantic category and its position in the corresponding list of words. Polysemous words should be represented in several groups since they have different meanings. Usually not more than two groups are needed for coarse-grained sense distinctions of these words. The meanings of words inside the linearly ordered groups also slightly differ, though these differences should be considered as fine-grained in contrast with coarse-grained sense distinctions between different groups.

We observe that semantic categories in the domain of positive meanings (in the negative and neutral as well) contain different number of words and the distribution of group sizes is quite regular. If arranged as parallel stripes all groups of positive adjectives fill the area nearly identical to a circular sector. The same holds for neutral and negative adjectives. The whole thesaurus of adjectives can be organized into the structure shown in Fig.3.

## 3 Discussion

We conjecture that the mapping onto the flat surface reflects representation of words in the patch of the cortex, rather than in an abstract space. Words (adjectives or their representations) occupy limited portion of the cortical surface and are distributed across this area in accordance with their semantic content. Three sectors of this cortical patch include words related to positive, neutral or negative quality, while stripes group together and order words related to a certain concept.

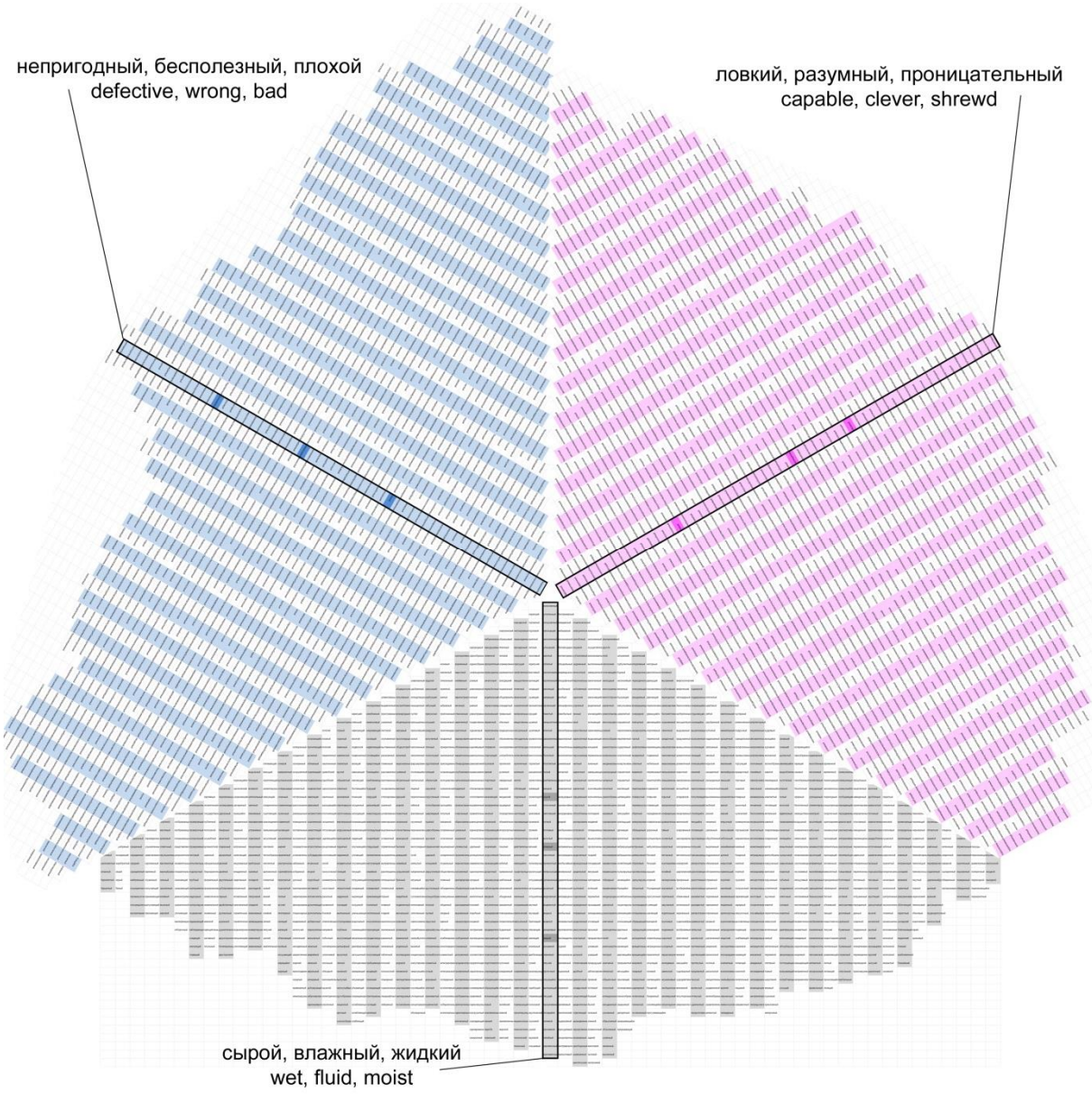


Fig.3. Complete set of linearly ordered groups of semantically close Russian adjectives, densely packed on the plane. Since there are about 6000 words, each one is written in the very small font. Three sectors of different colors contain words with neutral (gray), negative (blue) and positive (magenta) meanings. The groups, arranged in parallel stripes, are intermittently highlighted to distinguish adjacent ones. The largest semantic categories in each sector are shown in frames. Positions of important words inside each group are indicated in a darker color. These Russian words are listed in the callouts with English translations. They give the idea of the semantic content of each group.

Fig.3 shows congregations of neurons (could be cortical columns) which store neural representations of individual words. Each word occupies “personal” memory cell. The content of the cell can be used as a template for comparison with acoustic signals from ears during recognition of the spoken word. It is probable that just these memory cells generate sequences of neural spikes activating articulation muscles when the stored word is uttered. Some inferences about the properties of this cortical region can be drawn right now. First, there are many such areas in the cortex, since words in other lexical categories are organized in a similar

manner. We have preliminary results on verbs and adverbs. Bilinguals, most probably, use double set of such areas in their brain – one set for each language.

The human brain is quite proficient at word-sense disambiguation, which still remains a tough challenge for computer systems. The structure in the brain, similar to shown in Fig.3, can perform sense disambiguation automatically and in a highly efficient manner since humans do not feel polysemy to be a problem at all. Context preceding perception of a word "prepares" proper sector and the linear group inside, so that the acoustic input has to be compared with a set of words with close meaning. The word is anticipated. We believe that the final readout of the word from the memory is performed by a wave propagating from the center of the circular structure since we observe experimentally that the recognition time gradually grows for different words in ordered semantic category [7]. Travelling excitation waves in the human cortex were observed experimentally [13] and their function was analyzed theoretically [14].

The complex structure storing words should be studied in more detail both linguistically and experimentally, using neuroimaging techniques. This study should include experiments with children. The structure shown in Fig.3 is inborn containing no words for toddlers. The memory cells are initially empty and are filled with words in the process of language acquisition. Every new adjective has to find proper place in a certain quality domain and to squeeze in between words already in place. Monitoring the process of filling the area with words is a fascinating topic for experimental study.

## References


[1] Samsonovic A. V., Ascoli G. A., (2010). Principal Semantic Components of Language and the Measurement of Meaning. PLoS One. 2010; 5(6): e10921.

[2] Gärdenfors, P. (2014). Geometry of meaning: semantics based on conceptual spaces. Cambridge, Mass.: MIT Press. ISBN 9780262026789

[3] WordNet: An Electronic Lexical Database, available at: https://wordnet.princeton.edu

[4] Huth A. G., de Heer W. A., Griffiths T. L., Theunissen F. E., Gallant J. L. (2016). Natural speech reveals the semantic maps that tile human cerebral cortex. Nature, 532 (7600), 453–8. PMID: 27121839.

[5] Rodd J.M., Davis M.H. (2017) How to study spoken language understanding: a survey of neuroscientific methods. Language, Cognition and Neuroscience, 32:7, 805-817, DOI:10.1080/23273798.2017.1323110.

[6] Vvedensky V., Gurtovoy K., Sokolov M., Matveev M. (2020) Ordering of Words by the Spoken Word Recognition Time. In: Kryzhanovsky B., Dunin-Barkowski W., Redko V., Tiumentsev Y. (eds) Advances in Neural Computation, Machine Learning, and Cognitive Research III. NEUROINFORMATICS 2019. Studies in Computational Intelligence, vol 856. Springer, Cham (2020).

[7] Vvedensky V.L., Gurtovoy K.G. (2021). Topology of the Thesaurus of Russian Adjectives Revealed by Measurements of the Spoken Word Recognition Time. B. Kryzhanovsky et al. (Eds.): NEUROINFORMATICS 2020, SCI 925, pp. 1–6, 2021. https://doi.org/10.1007/978-3-030-60577-3_9

[8] Google Translate, available at: https://translate.google.com/

[9] Reverso Context , available at: http://context.reverso.net/

[10] Translate.academic.ru, available at: https://translate.academic.ru/

[11] Oxford Dictionaries, available at: https://en.oxforddictionaries.com/

[12] Ляшевская О. Н., Шаров С. А., Частотный словарь современного русского языка (на материалах Национального корпуса русского языка). М.: Азбуковник, (2009). http://dict.ruslang.ru/freq.php

[13] Martinet L.-E., Fiddyment G., Madsen J.R., Eskandar E.N., Truccolo W., Eden U.T., Cash S.S., Kramer M.A.. Human seizures couple across spatial scales through travelling wave dynamics. Nature Communications volume 8, Article number: 14896 (2017).

[14] Muller L., Chavane F., Reynolds J., Sejnowski T.J., Cortical travelling waves: mechanisms and computational principles. Nature Reviews Neuroscience 19 (5) pp 255-268. doi: 10.1038/nrn.2018.20 (2018).