



приоритет2030^
лидерами становятся



National Research
**Tomsk
State
University**

International Research & Practice Seminar

Experimental Studies of Language and Speech: Resource Bases and Technologies

supported by Tomsk State University Development Programme
(Priority 2030)

October 25, 2022

Experimental Studies of Language and Speech (E-SoLaS-2022)

**«Экспериментальные исследования языка и речи:
ресурсные базы и технологии»
(E-SoLaS-2022)**

**Проводится при поддержке Программы развития
Томского государственного университета
(Приоритет-2030)**

25 октября 2022 года

Hosts

Linguistic Anthropology Laboratory (<http://illa.tsu.ru/>)

Faculty of Philology, Tomsk State University, Russia (<http://www.tsu.ru/>)

Organizing Committee

Chair of Organizing Committee

Zoya I. Rezanova, Professor, Head of the Department of General, Computer and Cognitive Linguistics,
Deputy Head of the Linguistic Anthropology Laboratory

Co-chair of Organizing Committee

Olga V. Nagel, Dean of the Faculty of Foreign Languages; Senior Research Fellow, Linguistic Anthropology
Laboratory

Deputy Chair

Irina G. Temnikova, Associate Professor of the Department of English Philology, Faculty of Foreign
Languages; Senior Research Fellow, Linguistic Anthropology Laboratory

Andrei A. Stepanenko, Assistant of the Department of General, Computer and Cognitive Linguistics, Faculty
of Philology; Junior Research Fellow, Linguistic Anthropology Laboratory

Working group

Valeriia E. Vladimirova, postgraduate student of the Department of General, Computer and Cognitive
Linguistics, Faculty of Philology, Junior Research Fellow, Linguistic Anthropology Laboratory

Veronika S. Dibrova, postgraduate student of the Department of General, Computer and Cognitive
Linguistics, Faculty of Philology; Junior Research Fellow, Linguistic Anthropology Laboratory

Executive Secretary

Julia E. Sypchenkova, Research Assistant, Linguistic Anthropology Laboratory

Presentation formats:

Plenary Session Minutes: presentation (20 minutes) and discussion (10 minutes)

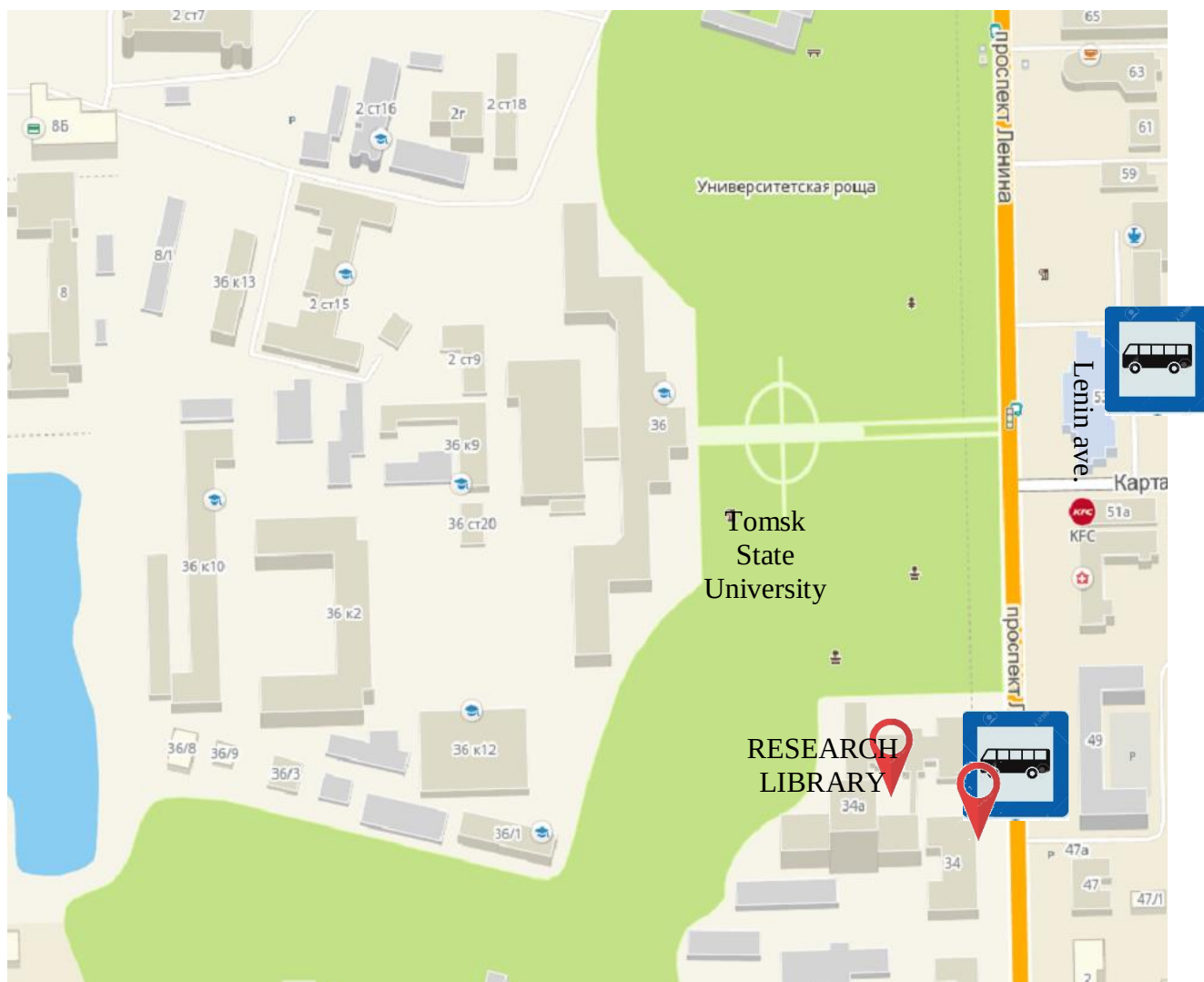
Paper Session Minutes: presentation (10 minutes) and discussion (5 minutes)

Working languages: Russian, English

Conference time: Tomsk (GMT+7)

Conference Venue

RESEARCH LIBRARY of Tomsk State University, Lenin Avenue, 34A, room 07



How to get there:

Bus stop Universitet

Buses: 2, 3, 4, 12, 19, 22, 23, 26, 31, 32, 33, 36, 37, 53, 112, 118, 442

Trolleybuses: 1, 3, 4

Bus stop Biblioteka TGU/TSU Library

Buses: 2, 3, 4, 12, 19, 22, 23, 26, 31, 32, 33, 36, 37, 53, 112, 118, 442

Trolleybus: 1, 3

Bus stop Yuridicheskiy Institut/Law Institute

Buses: 8, 9, 29

Programme overview

25.10.2022

RESEARCH LIBRARY of Tomsk State University, Room 07 / online (zoom) following the link
<https://us04web.zoom.us/j/3053592963?pwd=RnVuMIJPQ3hxKzRLajg4bjYxWEVsdz09>

Meeting ID: 305 359 2963

Passcode: 123

9.45 Registration and greeting

Plenary session

Moderators Zoya Rezanova, Olga Nagel

10.00 – 10.30 Elena Artemenko, Maksim Terpilovskii, Taisiia Ulianova, Victoria Vziatysheva, Olessia Koltsova

Predictors of message credibility in misinformation processing: eye-tracking studies

10.30 – 11.00 Mikhail Vlasov, Vlada Repeykova, Ulyana Trofimova

How to Measure Creative Thinking? From remote associates test to interpretation of phraseological units

11.00 – 11.30 Anastasiia Kolmogorova

The interface for measurement of quality and quantity of emotional component in text as a tool for psycholinguistic experiment

11-30 – 11-45 **Coffee break**

11-45 – 12-15 Oksana Tsaregorodtseva

Does language activate the sensorimotor properties of the entities it refers to and, if so, under what circumstances?

12-15 – 12-45 Andrei Stepanenko, Zoya Rezanova, Alexei Bocharov, Valery Datsuk, Dinara Aisheva

World2News service architecture: results and perspectives of multidisciplinary methods in the field of natural language processing

12-45 – 13 -15 Anna Sergina, Nikita Lysov, Nikita Mosin, Michail Denin, Natalia Litvinova

Features and functionality of the software product "Universal digital platform for integrating the functions of preparing, analyzing, storing and visualizing data for the tasks of cognitive linguistics"

13.15 – 14.30 **Lunch**

Paper Session

Moderators Elena Artemenko, Oksana Tsaregorodtseva

14.30 – 14.45 Kseniia Krapivina

Extracting and processing methods for user interface strings linguistic analysis

14-45 – 15-00 Ivan Pimenov

Annotating Argumentation in Texts Under Discursive Approach

15-00 – 15.15 Elena Pogodaeva

Thematic annotation in a RuTuBiC oral speech corpus: material specifics and annotation principles

15.15 – 15.30 Bella Kozlova

Sentiment analysis on the material of corpus of fiction synopsis

15.30 – 15.45 Dinara Aisheva

Text summarization with Transformer for the headline generation in Russian

15.45 – 16.00 Daria Belogorskaya

Development of criteria for evaluating news headlines based on the effectiveness of automatic summarizationalgorithms

16-00 – 16-15 **Coffee break**

16.15 – 16.30 Irina Korshunova

Collection of data and creation of a psycholinguistic database RuTurkPsychLing

16.30 – 16.45 Valeriia Vladimirova

Creation of a psycholinguistic database of native languages of heritage Turkic-Russian bilinguals (TurkWordPerception)

16.45 – 17.00 Umidjan Makhmudov

Unimodal and polymodal words in Uzbek and Russian

17.00 – 17.15 Veronika Dibrova

The influence of the education level on the subjective assessment of the language level of the heritage bilinguals (the Turkic-Russian languages' interaction of Southern Siberia)

17.15 – 17.30 Natalya Subbotina

Modern technology of collecting and processing language data within the experimental materials the Altai-Kizhi's sub-dialects

17.30 – 17.45 Victoria Zavarzina, Tatyana Litvinova

A database to study word associations in Russian with keystroke data

Программа семинара

25.10.2022

Научная Библиотека Томского государственного университета, 07 ауд. / онлайн по ссылке

<https://us04web.zoom.us/j/3053592963?pwd=RnVuMIJlPQ3hxKzRLajg4bjYxWEVsdz09>

Meeting ID: 305 359 2963

Passcode: 123

9.45 Регистрация и приветствие участникам

Пленарное заседание

Модераторы Зоя Резанова, Ольга Нагель

10.00 – 10.30 Елена Артеменко, Максим Терпиловский, Таисия Ульянова, Виктория Взятышева, Олеся Кольцова.

Предикторы доверия к новостным сообщения при обработке недостоверной информации: окуломоторное исследование

10.30 – 11.00 Михаил Власов, Влада Репейкова, Ульяна Трофимова

Как измерять креативное мышление? От теста отдаленных ассоциаций к интерпретации фразеологических единиц

11.00 – 11.30 Анастасия Колмогорова

Интерфейс для оценки качества и количества эмоции в тексте как инструмент психолингвистических исследований

11.30 – 11.45 Кофе-брейк

11.45 – 12.15 Оксана Царегородцева

Активирует ли язык сенсоримоторные свойства объектов и если да, то при каких обстоятельствах?

12.15 – 12.45 Андрей Степаненко, Зоя Резанова, Алексей Бочаров, Валерий Дацюк, Динара Аишева

Архитектура сервиса w2n: результаты и перспективы интеграции междисциплинарных методов в области обработки естественного языка

12.45 – 13.15 Анна Сергина, Никита Лысов, Никита Мосин, Михаил Денин, Наталия Литвинова

Функционал программного продукта «Универсальная цифровая платформа, интегрирующая функции подготовки, обработки, хранения и визуализации данных для задач когнитивной лингвистики»

13.15 – 14.30 **Обед**

Секционные доклады

Модераторы Елена Артеменко, Оксана Царегородцева

14.30 – 14.45 Ксения Крапивина

Методы извлечения и обработки материала для лингвистического анализа строк интерфейса пользователя

14.45 – 15.00 Иван Пименов

Аргументационное аннотирование текстов в рамках дискурсивного подхода

15.00 – 15.15 Елена Погодаева

Тематическая разметка в корпусе устной речи RuTuViC: специфика материала и принципы аннотирования

15.15 – 15.30 Белла Козлова

Анализ тональности на материале корпуса кратких содержаний художественных произведений

15.30 – 15.45 Динара Аишева

Суммаризация текста при помощи нейронной сети Transformer для генерации новостных заголовков

15.45 – 16.00 Дарья Белогорская

Разработка критериев оценки новостных заголовков на основе результативности алгоритмов автоматического реферирования

16.00 – 16.15 Кофе-брейк

16.15 – 16.30 Ирина Коршунова

Сбор материала и создание психолингвистической базы данных RuTurkPsychLing

16.30 – 16.45 Валерия Владимирова

Создание психолингвистической базы данных родных языков херитажных тюркско-русских билингов (TuWordPerception)

16.45 – 17.00 Умиджан Махмудов

Унимодальные и полимодальные слова в узбекском и русском языках

17.00 – 17.15 Вероника Диброва

Влияние уровня образования на субъективное оценивание уровня материнского языка херитажными билингвами (на материале тюркско-русского взаимодействия языков Южной Сибири)

17.15 – 17.30 Наталья Субботина

Современные технологии сбора и обработки языковых данных на основе экспериментального языкового материала говоров диалекта алтай-кижи

17.30 – 17.45 Виктория Заварзина, Татьяна Литвинова

База данных словесных ассоциаций, содержащая информацию о клавиатурном поведении респондентов

17-45 – 18.00 **Заключительное пленарное заседание**

Фуршет

Abstracts report

Plenary reports

Predictors of message credibility in misinformation processing: eye-tracking studies.

Elena Artemenko^{*1,2}, Maksim Terpilovskii¹, Taisiia Ulianova¹, Victoria Vziatysheva, Olessia Koltsova¹

¹ HSE University, Saint-Petersburg, Russia

² National Research Tomsk State University, Tomsk, Russia

*correspondence author: nekrasovaed@gmail.com

The spreading of misinformation in social media is now becoming apparent. In this study on news consumers' perceptions of message credibility we focus on two broad groups of factors: individuals' psychological and cognitive dispositions, and features of the message itself. The talk presents the results of two eye-tracking studies examining the effects of various predictors on fake news categorization accuracy and message credibility.

The first experiment is the text-level study, which examines the influence of various text elements, such as the source of the news message, on the message credibility and the consumer's ability to distinguish a false message from a true one. The second experiment examines the influence of cognitive bias features, specifically, confirmation bias. Participants' psychometric data (e.g., CRT) were also collected.

Each study is a laboratory experiment with the collection of eye movement data. For each experiment, the special interface was designed to mimic a social networking feed. Each news item contained a headline, the type of news source, the date of publication, and the text of the news item. Participants were asked to rate the credibility of each news item on a six-point Likert scale ranging from "true" to "false. Accuracy (correctness of news recognition, understood as percentage of hits) and credibility (measured on a six-point scale) were calculated as dependent variables. The independent variables in the two experimental studies are the source of the news, thinking style (as defined by CRT) and confirmation bias. Different results were obtained for different independent variables.

How to Measure Creative Thinking? From remote associates test to interpretation of phraseological units

Mikhail Vlasov^{*1}, Vlada Repeyko², Ulyana Trofimova³

¹Shukshin Altai State University for Humanities and Pedagogy, Biysk, Russia

²National Research University Higher School of Economics

³Russian Orthodox University, Moscow, Russia

Motivation

Creativity is assumed to comprise two cognitive processes. Convergent (concentrated) and divergent (dispersed) thinking play important roles in solving complex and non-standard problems (in engineering, business or research).

Problem statement

There is still lack of valid Russian-language tools to measure these processes. The elucidation of such linguistic abilities as formation of infrequent compound words or the ability to "metaphorize" can shed light on the essence of creativity.

We have proposed a new compound Remote Associates Test in Russian (Study 1) as well as another technique based on phraseological units' interpretation (Study 2).

Results

In the first study of compound RAT development, we found that both the task constrains and verbal intelligence of participants had significant effects on test performance. Moreover, we found one of the plausible psychophysiological implicit measures of divergent thinking process during RAT problem solving.

In the second study, we are trying to understand empirically why most of our Russian monolingual participants successfully recognize the meaning of a word-for-word translation of Mongolian phraseological unit such as "*Ukusit' yamochku ladoni* (Mongolian: *Algany honhorhoj hazah*)", while only few of them can correctly interpret the following unit: "*Vskormlennyj telenok telezhku razrushit* (Mongolian: *Tezheesen byaruu tereg evdekh*)".

The results are presented in the form of new ways of processing and interpreting such material according to the Item Response Theory.

Implications

Developing creativity in modern students can help them to cope with the future challenges of the society and create innovations for better quality of life.

The interface for measurement of the quality and quantity of emotional components in texts as a tool for psycholinguistic experiment

Anastasia Kolmogorova
Siberian Federal university
nastiakol@mail.ru

Initially, the proposed interface was elaborated for the purposes of emotional text annotation within a project in the field of sentiment analysis. As we didn't find any data base in Russian with emotional text annotation, we faced the problem of creating a key instrument to obtain the data.

Our statement is that our interface has much more heuristic potential than we initially thought: it could be used for a large set of tasks on the field of psycholinguistics and cognitive science.

While working with the interface, the informants are offered four scales, between the poles of which they could put a marker (a slider): Sadness – Enjoyment, Disgust – Anger, Shame – Excitement, Fear – Surprise. In the middle of the scale there is a zero emotion point. This design of scales was elaborated because the "Lövheim Cube" model of emotions these oppositions form its internal diagonals. After annotation conducted in this way, for each text and on each scale we obtain emotional values. Further, we can use them as raw data or, after having applied some mapping technics, transfer them to text emotional coordinates.

Using this interface we obtained a dataset of 4000 Internet-texts in Russian annotated by 2000 informants, on its basis we trained a range of ML-models and designed an analyzer of emotions in Internet-texts in Russian.

While doing this research, we have discovered that such a tool allows us to observe correlations between emotions and many other parameters in text perception and understanding. For example, our experiments with measurements correlations between values of emotional assessment and personal empathy level or with a level of language acquisition in bilinguals showed rather inspiring results.

In our report, we would like to share some suggestions for further use of the interface to do research in the field of psycholinguistics and cognitive studies.

Does language activate the sensorimotor properties of the entities it refers to and, if so, under what circumstances?

Oksana Tsaregorodtseva

University of Tübingen, Germany

Tomsk State University, Russia

correspondence author: caregrad@yandex.ru

There are many studies suggesting that at least part of our cognition, and especially our understanding of language, is embodied in actions, perceptions, and emotions and is therefore multimodal.

Whereas researchers in previous decades have mostly debated whether or not cognition is actually embodied, today the debate is more centered around the question of what circumstances or boundaries might limit sensorimotor activation if it occurs.

The purpose of this article is to show the role of multimodal representations in various tasks and contexts. In the first study, we will discuss the role of procedure and stimulus set. The second study will focus on the role of grammatical aspects in whether sensorimotor representations are activated or not. The latest study will demonstrate the effect of spatial cues on the activation of sensorimotor representations in a memory task.

World2News service architecture: results and perspectives of multidisciplinary methods in the field of natural language processing.

Zoya Rezanova¹, Andrei Stepanenko*¹, Alexei Bocharov¹, Valery Datsuk¹, Dinara Aisheva¹

¹ National Research Tomsk State University, Tomsk, Russia

*correspondence author: stepanenkone@mail.ru

An increase in the information flow and significant change in the realities of modern transforming society determines the work optimization of news agencies. Reputational risks have always been an integral part of every journalist's work. The world2news service allows you to effectively optimize the workforce of companies, automate part of the work by integrating interdisciplinary methods in the field of artificial intelligence.

The modularity of the system allows the different news agencies to effectively adapt to the needs of the customer, which include the following requests:

Automatic speech recognition module. The service includes the VOSK speech recognition system. The work is underway to collect sound texts for further fine-tuning of the model on the genre of media texts.

The module for extracting facts and quotes. Formal grammars have been developed using Tomita-parser software. Formal grammars based on similar rules allow summarizing texts.

The module for generating headlines and the news lead. The model includes trained models of seq2seq and ruGPT-3 neural networks.

The database module, which is the basis of the ontological knowledge base, includes basic statistical and linguistic data on news texts: tf-idf – statistics module, part of speech tagging.

Module for automatic classification of news texts (rubricator) using the Keras neural network

The data on which the models were trained had been collected from various sources, including RIA TASS, including regional news. The training, model setup, and development of formal grammars were carried out using the Python3 programming language. The formal accuracy metric used in the basis of generative models is ROUGE, the basis of the classifier is F1-score.

Features and functionality of the software product "Universal digital platform for integrating the functions of preparing, analyzing, storing and visualizing data for the tasks of cognitive linguistics"

Anna Sergina*¹, Nikita Lysov¹, Nikita Mosin¹, Michail Denin¹, Natalia Litvinova¹

¹Intelligent Profit Solutions Tomsk (IPST)

*correspondence author: sannaipstomsk@gmail.com

In recent years, cognitive linguistics has become one of the most dynamically developing areas of research, resulting in the need for modern digital tools for collecting and analyzing the data. The currently existing platforms designed to create and conduct research studies in such areas as cognitive linguistics, neurobiology, experimental psychology, etc., are mostly characterized by a complex interface with an unnecessary excessive functionality, the need be able to code in Python and other programming languages, as well as the necessity to use multiple software programs to create and conduct an experiment and analyze its results afterwards. These problems make the processes of creating and participating in the experiments significantly more complex and time-consuming.

The developing platform will help employees of research laboratories in the field of cognitive linguistics not only to create and publish their experiments, but to conduct them in the form convenient for the participants, collect and analyze responses, and generate statistics. The key benefits of this software product are the possibility to create experiments using a cross functional and user-friendly constructor without programming languages, as well as the convenience of interaction between the experiment creator and the respondents. In addition, the system simplifies the process of data collection and relieves the creators of experiments from the need to analyze the collected responses using external programs.

During the presentation the relevance and main goals of the project will be explained, the stages of product design and development presented, followed by the demonstration of the platform's interface and functionality.

Paper Session

Extracting and processing methods for user interface strings linguistic analysis

Kseniia Krapivina
Tomsk State University
krapivinaxena@gmail.com

One of the main goals in software development is to achieve the best usability possible, and among the tasks set is the prevention of any communication failures in user-system interaction. To reach that goal, clear understanding of interface inner working is necessary, including the verbal code nature, structure, role, function, and writing techniques ensuring its efficiency. The goal of current research is to explore textual and discourse specifics of user interface content. Hence we need to perform a comprehensive lexical, syntactical and semantic analysis as well as comparative analysis within the same or close domain on hundreds to thousands individual interface units («strings»). Moreover, for rhetorical analysis we need to formalize non-verbal relations between individual units as well as semantically linked groups. Thus, we need to establish timesaving and reproducible

methods of extraction, organizing and processing of aforesaid material. To accomplish that, we are developing the specialized markup tags system intended for use in subsequent automatic analysis.

Annotating Argumentation in Texts Under Discursive Approach

Ivan Pimenov
Novosibirsk State University
pimenov.1330@yandex.ru

The presentation focuses on the task of delineating the theoretical basis for annotating argumentation structures in texts. Such creation of argument-annotated corpora enables application of machine learning methods to automatic argumentation analysis (for instance, for computerized evaluation of texts persuasiveness, for identifying the reasons why specific points of view are supported). Delineation of the theoretical basis consists in comparison of the traditional approaches within the argumentation research, both from the linguistic perspective as well as in other fields.

The analyzed theories include Frans Eemeren's pragma-dialectical approach, Chaïm Perelman's rhetorical representation of argumentation, Jaakko's Hintikka's formal-logical approach, the Radical Argumentativism of Jean-Claude Anscombe and Oswald Ducrot, Marion Carel's theory of Semantic Blocks, and Ruth Amossy's discursive approach. The presentation then accentuates the three main points of Amossy's argumentation theory and demonstrates similarity of these points to the applied principles of argumentation modelling, defined in the Argument Interchange Format (AIF) standard.

To demonstrate the distinction of AIF as a specialized model for the formal argumentation representation, the presentation compares it with the generalized models of text organization (William Mann's Rhetorical Structure Theory, Segmented Discourse Representation Theory of Alex Lascarides and Nicholas Asher). This comparison consists in the elaboration of the specific traits of argumentation structures, which complicate their representation with the generalized models. The talk concludes with describing both the three core tasks in AIF-based argumentation annotation (identifying argumentative statements, establishing relations between them, specifying semantic models in relations) and the use of machine learning methods in addressing each.

Thematic annotation in the RuTuBiC oral speech corpus: material specifics and annotation principles

Pogodaeva Elena Nikolaevna
Tomsk State University
elenanikolaevnapogodaeva@gmail.com

The report describes the system of thematic annotation in a corpus of Russian oral non-public speech of Turkic-Russian bilinguals (RuTuBiC). So far, many attempts have been made to create universal recommendations for different annotation types, including thematic annotation. However, even if there are globally recognized standards, the final version of the annotation system depends on the purpose of creating the corpus and the characteristics of the language material. Also, the analysis of existing practices of corpus annotation showed the lack of systematic use of natural language processing methods for the implementation of corpus text metadata annotation. The main specificity of the analyzed textual material is the thematic heterogeneity of the dialogues with the genre-conditioned thematic constraints of the sociolinguistic interview. An experiment is described on the keyword extraction from manually annotated thematic fragments of subject semantics. It was conducted to assess its suitability in solving the problem of detecting topics in a dialogue corpus. Within this approach, each keyword is associated with weights which represent the importance of the words and are considered for the ranking. The results obtained can be used in adapting the classical topic modeling algorithm for topic detection in a dialogue.

Sentiment analysis on the material of corpus of fiction synopsis

Bella Kozlova

Higher School of Economics

bakozlova@edu.hse.ru

As a part of my current research I make content-based recommendation systems. In order to create the most personalized recommendations based on synopsis, I use different methods of processing language data: for example, automatic text processing and sentiment analysis.

The emphasis in the work is on sentiment analysis, which allows building and comparing plot schemes of different types of content. The use of synopses in this study follows from the assumption that synopses reflect the actual construction of the plot. Synopses also are convenient as material for research in case the original texts cannot be parsed or the dataset is too large.

As a corpus, I use a self-made dataset containing synopses of books, films, animated films and anime series. Using data scraping technologies I extracted main information about the works in dataset: the title and the author or director. The synopses were extracted manually. All further processing was fully automated and was carried out using the R and Python programming languages. The sentiment analysis was implemented both for each type of content separately and for the entire dataset.

The point of my research is in creating plot schemes using sentiment analysis, which will later be used in creation of recommendation systems.

Implications lie in the fact that the plot schemes were indeed identified using several indicators of the emotional coloring of the text. Now the received data can be successfully

used in the recommendation system. Also the collected corpus can serve as material for thematic modeling.

Text summarization with Transformer for headline generation in Russian

Dinara Aisheva
Tomsk State University
aishevadina@gmail.com

Due to increased use of neural networks, fine-tuning of big general language models is being used more often to save resources. One of the use cases of fine-tuned models is abstractive text summarization, the need for which has increased in recent years due to the rapid growth in the amount of information on the Internet. mT5 model is one of such big models, and it has been used to solve a number of problems; however, its applicability for news headline generation in Russian has yet not been researched profoundly. This research describes implementation of such a fine-tuned model for headline generation in Russian language, the analysis of the preliminary results in comparison with the simple version of a neural network architecture Seq2seq and the corpora used for fine-tuning. Despite being still not suitable for direct use, the results contribute to knowledge on the subjects of text summarization in Russian as well as the application of neural networks for solving problems of computational linguistics

Development of the criteria for evaluating news headlines based on the effectiveness of automatic summarization algorithms

Daria Belogorskaya
Tomsk State University
dariabelogorskaya@mail.ru

In recent years, the scientific community has faced the acute issue of solving the problem of increasing the volume of textual information, in particular, news texts. One of the ways to solve this problem is the use of algorithms for automatic texts summarization. The development of high-quality algorithms for automatic summarization requires the creation of effective systems for evaluating the effectiveness of these algorithms. One of the special cases of text summarization is the generation of news headlines, the evaluation of which is the focus of this study. Evaluation of the efficiency of the algorithm is carried out by philological reflection and consists in assessing the integrity of the headline, compliance with the syntactic norms of the Russian language, the degree of semantic identity of the original headline and the generated one. The material of the study was the

text data of the information agency "RIA Novosti" and "TASS" from the heading "Incidents". Headlines were generated using an algorithm developed in the world2news project. In the course of the analysis, the criteria for evaluating a news headline were determined and the primary typology of deviations that were identified in the analysis of the generated headlines was described. Identification of the types of deviations and the degree of regularity of their occurrence is an important first step towards improving the work of algorithms for automatic summarization of texts.

Collection of data and creation of a psycholinguistic database RuTurkPsychLing

Irina Korshunova
Tomsk State University
korshunova-61818@mail.ru

The report presents the creation of the psycholinguistic database and the description of the data collection steps for it. The creation of databases is an active trend in the world psycholinguistics. The materials presented in the database have an independent value, accumulating data on the perception and evaluation of stimuli by native speakers. At the same time, their main purpose is to be a resource base for the selection of representative stimulus material for subsequent psycholinguistic experiments studying cognitive processing of language. The RuTurkPsychLing PBD is being assembled by the questionnaire method. On the basis of assessments given by native speakers of the Russian language and Turkic-Russian bilinguals the contribution of psycholinguistic parameters to the semantics of Russian words is described. The presented database contains word assessments for the following parameters: familiarity, manipulability, location in space, correlation with size, emotionality, correlation with temperature. Respondents evaluated both Russian stimuli and stimuli in Turkic languages (Tatar, Khakass).

Creation of a psycholinguistic database of native languages of heritage Turkic-Russian bilinguals (TurkWordPerception)

Valeriia Vladimirova
Tomsk State University
picture_perfect@mail.ru

The report presents a psycholinguistic database of assessments of the contribution of five modalities of perception (taste, smell, touch, hearing and vision) to the semantics of the words of the Tatar and Khakass languages. The assessments were made by heritage

Turkic-Russian bilinguals. The creation of databases is an actively developing method in psycholinguistics. Databases provide an opportunity to model and evaluate the contribution of modal semantics on the basis of different groups of participants and languages with the ability to identify the specific features of the actualization of modal semantics. Databases can be also used to conduct contrastive studies between groups of participants. Moreover, databases provide ample opportunities for conducting psycholinguistic experiments, taking into account the contribution of perceptual modalities to the semantics of the evaluated words. Authors also present the formation of a vocabulary, the definition of calibrators in the process of adapting questionnaires for data collection, the content of base assessments, their structuring. The report will highlight research difficulties in creating such databases and ways to overcome them. As a result of the work, normative data on assessments of the contribution of modalities are presented, groups of respondents are compared, and their characteristic features are revealed. In general, the source study capabilities of the database in psycholinguistic research are characterized, as well as the results of the primary statistical analysis of the data, which makes it possible to identify certain patterns for each group.

Unimodal and polymodal words in Uzbek and Russian

Umidjan Makhmudov
Tomsk State University
abduvaliumid@yandex.com

Estimates of the role of sensory modalities in word meanings are currently being used as predictors in large mega-studies of word processing. This became possible because in recent years databases of sensory modality norms have been created for a number of languages: English, Dutch, French, Spanish, Chinese, Russian. As for the Turkic languages, as far as we know, psycholinguistic databases representing word scores on other psycholinguistically significant parameters exist only for the Turkish language, as well as small Turkic languages: Tatar, Khakass. Creation of a database of Uzbek words "UzRusWordPerception", the data of which can serve as a basis for conducting comparative studies on the material of other languages and, in general, will contribute to the introduction of the Uzbek language into the paradigm of psycholinguistic research.

The study used a questionnaire method using a 7-member Likert scale, as well as a comparison method. The data obtained will be proved by descriptive statistics and compared with the results of the RuWordPerception database.

When comparing the results, the following trends are observed: out of 60 Uzbek polymodal words, 29 words coincided with Russian polymodal words, many Uzbek bimodal

words coincided with Russian unimodal words. Uzbek translations of Russian words, which were rated below average in all modalities, almost coincided with Russian words.

The influence of the education level on the subjective assessment of the language level of the heritage bilinguals (the Turkic-Russian languages' interaction of Southern Siberia)

Veronika Dibrova
Tomsk State University
vorishka2012@gmail.com

At present, the Turkic languages of the small peoples of Southern Siberia are the languages of family communication. Therefore, sociolinguistic factors have a great influence on the language level. The study involved bilinguals whose native languages were Tatar, Khakass and Shor. The purpose of the study is to identify the degree of the influence of the education level on the assessment. The material was the evaluation of the results of a bilinguals' survey. The participants were divided into three groups: elementary education, secondary education (including vocational secondary and incomplete higher education) and higher education. The study used statistical analysis methods such as the Mann-Whitney U Test and One-way ANOVA analysis. The results indicate that there is no effect of education level on the language level of each of the three bilingual groups.

Modern technology of collecting and processing language data within the experimental materials the Altai-Kizhi's sub-dialects

Natalya Subbotina
Tomsk State University
subbotina1736@gmail.com

The expected presentation is based on the results of the processed experimental language material collected as a part of the field study in the territory of the Altai Republic. The study is conducted on the material of sub-dialects of the Altai-Kizhi dialect. The technologies and methods of collecting material allow us to study the vocalism of the southern dialects of the Altai language more qualitatively.

The study uses audio materials collected on sound recording equipment applying a pre-prepared language questionnaire. Computer programs necessary for the analysis and processing of the collected language material are also used.

The proposed results are aimed at obtaining an accurate description of the vocal system of each studied sub-dialect of the Altai-Kizhi dialect. Moreover, with the help of

acoustic and articulatory phonetic studies, it is necessary to find out the phonological features of the southern sub-dialects of the Altai language.

The results obtained will allow us to reliably establish the linguistic features of the southern sub-dialects of the Altai language and their areal distribution. Conclusions supported by the data of articulatory and acoustic-phonetic studies will be presented. Classifications of sub-dialects of the Altai-Kizhi dialect will also be established on the basis of accurate phonetic data using the comparative method. The research work will deepen and expand understanding in the study of Turkic languages, as well as provide more information for further study of sub-dialects of the Altai-Kizhi dialect.

A database to study word associations in Russian with keystroke data

Victoria Zavarzina*¹, Tatyana Litvinova¹

¹Voronezh State Pedagogical University

*correspondence author: zva0604@yandex.ru

Artificial intelligence (AI) have opened up new possibilities for the study of mental lexicon through word associations. Currently, there are various programs for recording keystrokes that can be used for this kind of research.

We offer a new look at word associations using the keystroke logging method. A special corpus was compiled from more than 12,000 associations for 50 cue words represented by different parts of speech. We collected associations from 49 adult participants studying at the university in various educational programs. Each participant gave at least 5 associations for each of the cue words. The experiment was undertaken at several stages which allows taking a close look into the stability of associations.

Based on the data from keystroke logging software (namely the duration of pauses between associations and the speed of typing of particular word), we can study the strength of associations.

In addition to associations, our database contains the information about respondents (personality test scores as well as results of emotional state test), which allows examining the possible effect of emotional state and personality traits on the stability, strength and type of associations.

Along with associations, our database contains texts on the topics related to the several cue words, which allows comparing associations of the cue words and their real use in the text. The texts are also accompanied by keystroke logging information.

To the best of our knowledge, this database is the first one to study word associations from different perspectives including keystroke data.

This work was supported by Russian Science Foundation, project number 21-78-10148 “Modeling the meaning of a word in individual linguistic consciousness based on distributive semantics”.

Our members

Aisheva Dinara, Tomsk State University
Artemenko Elena, Higher School of Economics
Belogorskaya Daria, Tomsk State University
Bocharov Alexei, Tomsk State University
Datsuk Valery, Tomsk State University
Denin Michail, Intelligent Profit Solutions Tomsk (IPST)
Dibrova Veronika, Tomsk State University
Koltsova Olessia, Higher School of Economics
Kolmogorova Anastasia, Siberian Federal university
Korshunova Irina, Tomsk State University
Kozlova Bella, Higher School of Economics
Krapivina Kseniia, Tomsk State University
Litvinova Natalia, Intelligent Profit Solutions Tomsk (IPST)
Litvinova Tatyana, Voronezh State Pedagogical University
Lysov Nikita, Intelligent Profit Solutions Tomsk (IPST)
Makhmudov Umidjan, Tomsk State University
Mosin Nikita, Intelligent Profit Solutions Tomsk (IPST)
Pimenov Ivan, Novosibirsk State University
Repeykova Vlada, Higher School of Economics
Rezanova Zoya, Tomsk State University
Sergina Anna, Intelligent Profit Solutions Tomsk (IPST)
Stepanenko Andrei, Tomsk State University
Subbotina Natalya Tomsk State University
Terpilovskii Maksim, Higher School of Economics
Trofimova Ulyana, Russian Orthodox University
Tsaregorodtseva Oksana, Tomsk State University
Ulianova Taisiia, Higher School of Economics
Vlasov Mikhail, Shukshin Altai State University for Humanities and Pedagogy
Vziatysheva Victoria, Higher School of Economics
Zavarzina Victoria, Voronezh State Pedagogical University

Наши участники

Аишева Динара Армановна, аспирант кафедры общей, компьютерной и когнитивной лингвистики, младший научный сотрудник лаборатории лингвистической антропологии ТГУ, Томск

Артеменко Елена Дмитриевна, кандидат филологических наук, сотрудник лаборатории социальной и когнитивной информатики НИУ ВШЭ, Санкт-Петербург; старший научный сотрудник лаборатории лингвистической антропологии филологического факультета НИ ТГУ, Томск

Белогорская Дарья Викторовна, аспирант кафедры общей, компьютерной и когнитивной лингвистики НИ ТГУ, Томск

Бочаров Алексей Владимирович, кандидат исторических наук, доцент кафедры истории древнего мира, средних веков и методологии истории НИ ТГУ, Томск

Взятышева Виктория Сергеевна, стажер-исследователь лаборатории социальной и когнитивной информатики Национального исследовательского университета «Высшая школа экономики», Санкт-Петербург

Владимирова Валерия Евгеньевна, аспирант кафедры общей, компьютерной и когнитивной лингвистики; младший научный сотрудник лаборатории лингвистической антропологии НИ ТГУ, Томск

Власов Михаил Сергеевич, кандидат филологических наук, доцент, Алтайский государственный гуманитарно-педагогический университет имени В. М. Шукшина, Бийск

Дацюк Валерий Валентинович, старший преподаватель кафедры общей, компьютерной и когнитивной лингвистики; младший научный сотрудник лаборатории лингвистической антропологии НИ ТГУ, Томск

Денин Михаил, Intelligent Profit Solutions Tomsk (IPST)

Диброва Вероника Сергеевна, аспирант кафедры общей, компьютерной и когнитивной лингвистики; младший научный сотрудник лаборатории лингвистической антропологии НИ ТГУ, Томск

Заварзина Виктория Александровна, аспирант, лаборант научной лаборатории компьютерной семасиологии ВГПУ, Воронеж

Козлова Белла Алексеевна, студент НИУ ВШЭ, Санкт-Петербург

Колмогорова Анастасия Владимировна, доктор филологических наук, профессор, заведующая кафедрой романских языков и прикладной лингвистики ИФ и ЯК Сибирского федерального университета, Красноярск

Кольцова Олеся Юрьевна — кандидат социологических наук, заведующая Интернет-исследований НИУ ВШЭ, доцент факультета социологии НИУ ВШЭ, Санкт-Петербург

Коршунова Ирина Сергеевна, аспирант кафедры общей, компьютерной и когнитивной лингвистики; младший научный сотрудник лаборатории лингвистической антропологии НИ ТГУ, Томск

Крапивина Ксения Петровна, аспирант кафедры общей, компьютерной и когнитивной лингвистики НИ ТГУ, Томск

Литвинова Наталия, Intelligent Profit Solutions Tomsk (IPST)

Литвинова Татьяна Александровна, доктор филологических наук, научный сотрудник регионального центра русского языка; ведущий научный сотрудник научно-исследовательской лаборатории компьютерной семасиологии, ВГПУ, Воронеж

Лысов Никита, Intelligent Profit Solutions Tomsk (IPST)

Махмудов Умиджан Реимбаевич, аспирант кафедры общей, компьютерной и когнитивной лингвистики НИ ТГУ, Томск

Мосин Никита, Intelligent Profit Solutions Tomsk (IPST)

Пименов Иван Сергеевич, аспирант НИ НГУ, Новосибирск

Погодаева Елена Николаевна, аспирант кафедры общей, компьютерной и когнитивной лингвистики НИ ТГУ, Томск

Резанова Зоя Ивановна, доктор филологических наук, профессор, заведующая кафедрой общей, компьютерной и когнитивной лингвистики, зам. зав. лаборатории лингвистической антропологии НИ ТГУ, Томск

Репейкова Влада Алексеевна, ? НИУ ВШЭ, Москва

Сергина Анна, Intelligent Profit Solutions Tomsk (IPST)

Степаненко Андрей Александрович, младший научный сотрудник лаборатории лингвистической антропологии НИ ТГУ, Томск

Субботина Наталья Викторовна, аспирант кафедры общей, компьютерной и когнитивной лингвистики; младший научный сотрудник лаборатории лингвистической антропологии НИ ТГУ, Томск

Терпиловский Максим Александрович, кандидат биологических наук, старший научный сотрудник лаборатории социальной и когнитивной информатики НИУ ВШЭ, Санкт-Петербург

Трофимова Ульяна Михайловна, ? Российский православный университет святого Иоанна Богослова, Москва

Ульянова Таисия Андреевна, стажер-исследователь лаборатории социальной и когнитивной информатики НИУ ВШЭ, Санкт-Петербург

Царегородцева Оксана Викторовна, сотрудник лаборатории 'Язык и познание' Тюбингенского университета, старший научный сотрудник лаборатории лингвистической антропологии НИ ТГУ, Томск