Polina Eismont · Maria Khokhlova ·
Mikhail Koryshev ·
Elena Riekhakaynen *Editors*

Literature, Language and Computing

Russian Contribution

Springer

Literature, Language and Computing

Polina Eismont · Maria Khokhlova ·
Mikhail Koryshev · Elena Riekhakaynen
Editors

Literature, Language and Computing

Russian Contribution

Springer

Editors

Polina Eismont
St. Petersburg University
St. Petersburg, Russia

Maria Khokhlova
St. Petersburg University
St. Petersburg, Russia

Mikhail Koryshev
St. Petersburg University
St. Petersburg, Russia

Elena Riekhakaynen
St. Petersburg University
St. Petersburg, Russia

ISBN 978-981-99-3603-8 ISBN 978-981-99-3604-5 (eBook)
https://doi.org/10.1007/978-981-99-3604-5

This Springer imprint is published by the registered company Springer Nature Singapore Pte Ltd.
The registered company address is: 152 Beach Road, #21-01/04 Gateway East, Singapore 189721, Singapore

Preface

This volume presents the collection of papers presented at the international scientific conference “Literature, Language and Computer Technologies”, which was held at St. Petersburg University from November 10 to 12, 2022. The purpose of the event was to provide an opportunity for scientific dialogue between the specialists in the Russian language and Russian literature, as well as in comparative studies on the benefits of using computer technologies in their studies. The opportunities offered by modern computer technologies and formal methods (for example, corpus and quantitative analysis) led both to the emergence of new applied areas and to the manifestation of interest in existing ones and giving impetus to their development. The plenary reports were devoted to the issues of machine translation, the study of poetic texts and digitalization in philology. Research in the field of Russian studies became a separate iconic topic: the study of modern oral discourse, the resolution of anaphora and ambiguity in legislative texts, the analysis of literary texts, the creation of a multimedia corpus of ironic speech for phonetic studies, the development of companion robots, the study of Russian as a foreign language. The conference was held for the first time, but the organizers are sure that it will become regular and will attract more interested participants in the future, acquainting listeners with a wide range of problems that are successfully solved by linguists and literary critics using computer methods.

The papers included in the collection are divided into two parts: the first part contains articles based on the material of corpus research, and the second part presents articles that focus on the use of computer technology in linguistic research. The collection opens with an article by **Kristina Zaides**, **Daria Gorbunova and Natalia Bogdanova-Beglarian**, in which they discuss some crucial issue for the understanding of the process of speech production. The authors have included the factor of speaker psychological profile as a significant point and provided a reliable analysis of its influence on the self-correction strategies. The issues of syntactic organization of the text based on the data of written language corpora are solved in the articles by **Alexandra Chaga**, who discusses a Russian construction *den’-den’skoj* (≈ ‘day-to-day’) both in its formal and interpretive aspects using the microsyntactic approach as its framework, and by **Olga Blinova**, who presents the results of an analysis of the use of pronominal anaphora in legislative texts, which can be used in the future in solving the problem of the intelligibility of the official style of the Russian language for lawyers and other native speakers. The question of criteria for selecting texts when creating a historical corpus, the choice of the type of description, the structure of the corpus and the possibilities of using the developed corpus in teaching process at the university are the focus of the article by **Ilia Afanasev and Andrey Babanov**.

The collection continues a series of articles that consider data from other languages, but whose results are important and interesting for the development of translation activities. Thus, the article by **Yuliya Auseichyk** based on the Frantext corpus data and the comparative analysis of relative frequencies of some core French coordination units shows the relevance of varying macro- and microscales in a diachronic perspective and invite researchers to be more careful in handling data from corpora. **Ekaterina Ivanova, Olga Khutoretskaya, Maria Solovieva** rely on corpus-based procedures (11th–13th century texts from Frantext corpus) to examine the grammaticalization of *Futur Antérieur* and show that its observed incompleteness could be considered as a prerequisite for the development of this function. The last two articles of the first section demonstrate an example of using corpus and semantic methods when analyzing changes in the use of two German lexemes *Nachhaltigkeit/nachhaltig* (in the article by **Irina Jesan, Elena Kovtunova, Elena Sadovskaya**) and when accessing corpus data to verify Russian-English translation correspondences (in the article by **Adelya Abdulmanova, Ekaterina Vyunova, Irina Lekomtseva**).

The first part is completed by a block of three articles that represent literary corpus studies. **Aleksander Grebennikov, Ekaterina Ivanova, Mikhail Koryshev, Maria Solovieva** present a sample of a comparative computational study of translating Nabokov's critics into 4 European languages and claim that the suggested method of computer stylometric analysis makes it possible to trace the formation of new associative fields in the translated text. The article by **Alexander Pipersky** continues the topic of computational text analysis and, using the example of Russian poetic texts, presents a critical review of various quantitative methods for measuring lexical diversity. The article by **Marina Ponomareva** is devoted to the creation, organization, tagging and structure of the corpus of Russian literary texts of the 18th century.

The part two of the collection begins with the articles discussing the issues whose relevance has been undeniably proven by the recent pandemic—all of them consider the use of computer technology in linguistic education. **Elena Laskareva and Alina Pozdnyakova** focus on one of the most difficult topics in learning Russian—the Russian verb system—and offer a new structure of a digital dictionary that is aimed specifically at foreigners learning Russian as a foreign language. The use of a multilingual dictionary of East Slavonic proverbs in teaching language, translation and cultural linguistics is considered in the article by **Olga Raina, Viktoria Muschinskaya, Olga Guseva**. Articles by **Natalia Kucherenko, Tatiana Alexeytseva, Maria Miretina, Olga Voicou** and **Olga Antciferova, Tatyana Kolosova, Kira Shchukina** consider ways to combine offline and online methods in teaching a foreign language at a university—from integration of the informal learning of a foreign language with formal teaching methods to improve the benefits of unstructured self-education strategy to the adaptation of peer-to-peer teaching of Russian as a foreign language to the current distance teaching format. An interdisciplinary article by **Anna Tiraspolskaya** covers the issue of using the Internet forums and blogs as a source of texts in teaching language for special purposes, and based on her own experience she shows the benefits of this method. The last article of the section by **Natalia Shutemova** considers the actual problem of machine translation of artionyms that play a significant role in forming national and world artionymicon and participate in creating art discourse and proposes the ways how to improve the neural network machine translation systems in this sphere.

The collection is completed by three articles that consider the issues of representation and recognition of emotions in automatic processes of speech analysis and synthesis. **Anastasia Kolmogorova** and **Alexander Kalinin**'s article proposes the mechanisms of the visualization of emotions in text analysis. A multimedia annotated corpus of Russian texts with ironic statements, which is discussed in the article by **Uliana Kochetkova, Pavel Skrelin, Vera Evdokimova, Tatiana Kachkovskaia**, adequately compensates for the shortcomings in the availability of phonetically and prosodic annotated corpora available for research to solve pragmatic problems. The last article of the collection by **Artemy Kotov** and **Anna Zinina** discusses the development of new components of the F-2 emotional robot, the functionality of which provides for the possibility of reasoning and imagination that allows the robot to process statements that are not relevant to the current extralinguistic situation but are emotionally connected with it.

This volume will be a useful reference for scholars in theoretical and applied linguistics, literary studies, language teaching and education.

St. Petersburg, Russia

Polina Eismont
Mikhail Koryshev
Maria Khokhlova
Elena Riekhakaynen

Contents

Corpora in…… Language and Translation Studies

Self-Repair in Russian Spoken Discourse in Psycholinguistics Aspect: Correlation Analysis and Quantitative Data 3
Kristina Zaides, Daria Gorbunova, and Natalia Bogdanova-Beglarian

Den' Den'skoj: A Lexicographic Portrait of a Russian Microsyntactic Unit 21
Alexandra Chaga

"Plain and Natural" Versus "Accurate and Unambiguous": Pronominal Intrasentential Anaphora in Russian Legislative Texts 31
Olga Blinova

The Old Church Slavonic Corpora and Their Use in Language Studies at the University 43
Ilia Afanasev and Andrey Babanov

Core Coordination Units in Macro- and Microdiachrony: Experimental Data 59
Yuliya Auseichyk

The Use of *Futur Antérieur* in the Past in Old French: A Corpus-Based Study 69
Ekaterina Ivanova, Olga Khutoretskaya, and Maria Solovieva

Nachhaltigkeit in Media Crisis Discourses 81
Irina Jesan, Elena Kovtunova, and Elena Sadovskaya

Using Corpora for Verifying Language Choices in Translation 93
Adelya Kh. Abdulmanova, Ekaterina K. Vyunova, and Irina A. Lekomtseva

Corpora in…… Literature Studies

Stylometric Methods in Comparative Text Analysis 103
Alexander Grebennikov, Ekaterina Ivanova, Mikhail Koryshev, and Maria Solovieva

Lexical Diversity of Russian Poets 113
Alexander Piperski

A Semantic Corpus of Russian Literature of 18 Century: Its Current State and Its Future 121
Marina Ponomareva

Computing in…… Language Teaching and Translation

Multimedia Dictionary of Verbal Vocabulary: Concept, Structure, Implementation 131
Elena R. Laskareva and Alina A. Pozdnyakova

East Slavonic Proverbs of the "Learning—Inattention" Thematic Group (Based on the New Electronic Dictionary of Modern East Slavonic Proverbs) 143
Olga V. Raina, Viktoria V. Mushchinskaya, and Olga V. Guseva

Online Peer-To-Peer Teaching Practice as a Tool to Advance Professional Competence in Distant Learning Format 155
Olga Antciferova, Tatyana Kolosova, and Kira Shchukina

Incorporating Informal e-Learning into Foreign Language Teaching Through Collaborative Personalization 173
Natalia Kucherenko, Tatiana Alexeytseva, Maria Miretina, and Olga Voicou

Dental Internet Resources in Teaching Specialty Language Within Russian as a Foreign Language Course 185
Anna Yu. Tiraspolskaya

Machine Translation Versus Human Translation of Artionyms 197
Natalia Shutemova

Computing in…… Emotional Text Processing

The Emotion in Text Analyzer: How to Visualize Its Output 211
Anastasia Kolmogorova and Alexander Kalinin

The Multimedia Corpus of Russian Ironic Speech for Phonetic Analysis 223
Uliana Kochetkova, Pavel Skrelin, Vera Evdokimova, and Tatiana Kachkovskaia

Theory of Mind and the Mechanism of Imagination for a Companion Robot 239
Artemiy Kotov and Anna Zinina

Editors and Contributors

About the Editors

Polina Eismont is an Associate Professor at Ludmila Verbitskaya Department of General Linguistics, St. Petersburg University. Her research interests include Language Acquisition and Psycholinguistics, Text Linguistics, Cognitive Linguistics, Event Structure, Music Semantics, Syntax of Nulls. She obtained her Ph.D. from St. Petersburg University in 2008. She is the author of more than 50 papers in domestic and international journals and volumes and the co-editor of two CCIS volumes "Language, Music, and Computing" (Springer Verlag, 2015, 2019) and a book "Language, Music and Gesture: Informational Crossroads (LMGIC 2021)" (Springer Singapore, 2021).

Maria Khokhlova is an Associate Professor at Department of Mathematical Linguistics, St. Petersburg University. Her research lies at the intersection of natural language processing, corpus linguistics and machine learning and was supported by grants and scholarships awarded by Russian and international foundations (RSF, DAAD, Visegrad Fund, Erasmus). She holds a Ph.D. from St. Petersburg University (2011) and was awarded the St. Petersburg Government Prize for science and teaching (2020, 2022). She is the author of more than 100 articles in domestic and international refereed journals and volumes.

Mikhail Koryshev is an Associate Professor at Department of Comparative Studies of Languages and Cultures and the Dean of the Faculty of Philology, St. Petersburg University. His research interests focus on German language and culture, literary studies, as well as on Catholic hymnography and liturgiology. He obtained a Ph.D. in Philology from St. Petersburg University in 2005. He was a visiting lecturer in Germany, his research was supported by Erasmus and DAAD-Stiftung. He is the author of more than 60 papers in Russian and international peer-reviewed journals and book series.

Elena Riekhakaynen is an Associate Professor and the Head of Ludmila Verbitskaya Department of General Linguistics, St. Petersburg University. She obtained her Ph.D. from St. Petersburg University in 2011. Her research interests include Phonetics and Psycholinguistics, especially Spoken Word Recognition and Multimodal Text Processing. She is one of the developers of the Russian Speech Corpus and the author of three books and more than 60 papers in linguistics and interdisciplinary studies. For her research, she has been awarded prizes from St. Petersburg University, St. Petersburg City Administration, Russian Educational Foundations, and the Program Committees of several international conferences.

Contributors

Adelya Kh. Abdulmanova St. Petersburg University, St. Petersburg, Russia

Ilia Afanasev National Research University Higher School of Economics, Moscow, Russia;
MTS Artificial Intelligence Center, LLC, Moscow, Russia

Tatiana Alexeytseva St. Petersburg University, St. Petersburg, Russia

Olga Antciferova St. Petersburg University, St. Petersburg, Russia

Yuliya Auseichyk Minsk State Linguistic University, Minsk, Republic of Belarus

Andrey Babanov St. Petersburg University, St. Petersburg, Russia

Olga Blinova St. Petersburg University, St. Petersburg, Russia;
National Research University Higher School of Economics, St. Petersburg, Russia

Natalia Bogdanova-Beglarian St. Petersburg University, St. Petersburg, Russia

Alexandra Chaga Institute for Information Transmission Problems, Russian Academy of Sciences, Moscow, Russia

Vera Evdokimova St. Petersburg University, St. Petersburg, Russia

Daria Gorbunova St. Petersburg University, St. Petersburg, Russia

Alexander Grebennikov St. Petersburg University, St. Petersburg, Russia

Olga V. Guseva St. Petersburg University, St.Petersburg, Russia

Ekaterina Ivanova St. Petersburg University, St. Petersburg, Russia

Irina Jesan St. Petersburg University, St. Petersburg, Russia

Tatiana Kachkovskaia St. Petersburg University, St. Petersburg, Russia

Alexander Kalinin National Research University Higher School of Economics, St. Petersburg, Russian Federation

Olga Khutoretskaya St. Petersburg University, St. Petersburg, Russia

Uliana Kochetkova St. Petersburg University, St. Petersburg, Russia

Anastasia Kolmogorova National Research University Higher School of Economics, St. Petersburg, Russian Federation

Tatyana Kolosova St. Petersburg University, St. Petersburg, Russia

Mikhail Koryshev St. Petersburg University, St. Petersburg, Russia

Artemiy Kotov Kurchatov Institute, Moscow, Russia;
Russian State University for the Humanities, Moscow, Russia;
Moscow State Linguistic University, Moscow, Russia

Elena Kovtunova St. Petersburg University, St. Petersburg, Russia

Natalia Kucherenko St. Petersburg University, St. Petersburg, Russia

Elena R. Laskareva St. Petersburg University, St. Petersburg, Russia

Irina A. Lekomtseva St. Petersburg University, St. Petersburg, Russia

Maria Miretina St. Petersburg University, St. Petersburg, Russia

Viktoria V. Mushchinskaya St. Petersburg University, St.Petersburg, Russia

Alexander Piperski Russian State University for the Humanities, Moscow, Russia

Marina Ponomareva St. Petersburg University, St. Petersburg, Russia

Alina A. Pozdnyakova The Kosygin State University of Russia, Moscow, Russia

Olga V. Raina St. Petersburg University, St.Petersburg, Russia

Elena Sadovskaya St. Petersburg University, St. Petersburg, Russia

Kira Shchukina St. Petersburg University, St. Petersburg, Russia

Natalia Shutemova St. Petersburg University, St. Petersburg, Russia

Pavel Skrelin St. Petersburg University, St. Petersburg, Russia

Maria Solovieva St. Petersburg University, St. Petersburg, Russia

Anna Yu. Tiraspolskaya St. Petersburg University, St. Petersburg, Russia

Olga Voicou St. Petersburg University, St. Petersburg, Russia

Ekaterina K. Vyunova St. Petersburg University, St. Petersburg, Russia

Kristina Zaides National Research University Higher School of Economics, St. Petersburg, Russia

Anna Zinina Kurchatov Institute, Moscow, Russia;
Russian State University for the Humanities, Moscow, Russia;
Moscow State Linguistic University, Moscow, Russia

Corpora in…… Language and Translation Studies

Self-Repair in Russian Spoken Discourse in Psycholinguistics Aspect: Correlation Analysis and Quantitative Data

Kristina Zaides, Daria Gorbunova, and Natalia Bogdanova-Beglarian

Abstract The article is aimed at revealing and describing the correlation between the number and usage of self-repairs in monologues and the psychological characteristics of the speaker. Self-repair is the correction of one's own speech fragment which has already been produced. The usage of a particular self-repair strategy by different psychological groups of speakers can show how the speakers' psychotype can affect their speech production. The data for the study are 24 monologues-descriptions in Russian from the corpus "Balanced Annotated Text Library": 12 monologues of native Russian speakers and 12 monologues of foreign Russian speakers. In the monologues-descriptions of comic strips by H. Bidstrup "Hair loss treatment", the cases of self-repair were annotated dividing into error correction and self-editing. The main repair operation types were revealed: word form change, word change, replacement of a word part with a word, word insertion, and phrase reformulation. Lexical or non-lexical repair initiators were also tagged—breaks, hesitation pauses, vocalized pauses, prolongations of sounds, discourse markers, etc. The study shows that self-repairs in general are more frequent in introverts' speech. The prevalent type among all the speakers is self-editing, however, error correction is used more often by introverts. The most frequent repair operations in introverts' speech are replacement of a word part with a word and word form change. Extraverts' most frequent type of repair operations is word change. Introverts usually initiate repairs with breaks, silent and vocalized pauses, extraverts—more often with silent pauses, ambiverts—with breaks and silent pauses. Thus, self-repair initiation and type of operation is, in some cases, connected with the psychological characteristics of the speaker.

K. Zaides
National Research University Higher School of Economics, Soyuza Pechatnikov Str. 16, 190121 St. Petersburg, Russia
e-mail: kzaydes@hse.ru

D. Gorbunova (✉) · N. Bogdanova-Beglarian
St. Petersburg University, Universitetskaya Emb. 7/9, 199034 St. Petersburg, Russia
e-mail: dgorbunova2@gmail.com

N. Bogdanova-Beglarian
e-mail: n.bogdanova@spbu.ru

P. Eismont et al. (eds.), *Literature, Language and Computing*,
https://doi.org/10.1007/978-981-99-3604-5_1

Keywords Self-repair · Error correction · Self-editing · Hesitation pause · Break · Spoken discourse · Spontaneous speech · Monologue · Psycholinguistics

1 Introduction

Spoken discourse represents one of the forms of language existence full of different types of speech disfluencies—hesitation pauses, meta-communication, breaks, slips of tongue, and, as well, self-repair (or self-correction) [1]. There are plenty of genres of spoken speech, e.g., informal conversation, public talk, phone dialogue, etc. Monologue is one of the forms of spoken discourse which is considered less spontaneous than dialogic forms; however, monologues as well demonstrate all essential characteristics of spoken speech mentioned above. A picture description is a common type of monologues for field linguistics, when an informant describes a given picture and produces a relatively large spoken text—a monologue-description. A produced monologue-description "on the one hand, <…> is largely motivated by the image and constrained by the topic, which entails the use of a special kind of constructions and the corresponding vocabulary. On this basis, all monologues, which are descriptions of one image, can be considered isomorphic and can be compared with each other at different levels. On the other hand, a monologue-description, being a form of unprepared speech, has a sufficiently high degree of spontaneity and a set of stable features typical of spontaneous speech <…>" [2: 354]. Comparison of such monologues with each other, taking into account the psychological characteristics of the speakers, allows to reveal the correlation between the usage of some particular features inherent to spontaneous speech and speaker's personality and can bring an interesting result.

Self-repair is usually considered one of the characteristic phenomena of spontaneous speech alongside the hesitation pauses, breaks, prolongations, etc., which is implemented with the usage of this means, too. Self-repair is understood as a speaker's verbalized attempt to correct the part of the spoken discourse which has already been produced. It can be the correction of a particular error which has been noticed by the speaker, or editing of some speech fragment which was not satisfactory for the speaker:

*i potom on spal/i-i () i vdrug on () **i-i kogda on***[1] *prosnulsya/on-n () zametil chto/ u nego/tochno/uzhe yest' volos*[2] (S3, L2, w., 24, intr.)[3]—self-editing;

*vstayot/ryadom/**pered zerkalom/ryadom s rakovinoj**/n–no vidimo v vanne* (S1, L1, w., 20, extr.)—self-editing;

*Aleksandr nemnogo () ispugalas' **() ispugalsya*** (S7, L2, w., 27, intr.)—right error correction by non-native speaker;

*i v poslednem **() poslednej** risunki/my z… () znaem chto/etot muzhchina/(n-n) () vsyo () vsyo proiskhodit v yego sne* (S5, L2, m., 27, amb.)—wrong error correction by non-native speaker.

Any spontaneous text contains self-repairs, but, in monologues, self-repairs are especially interesting since they reflect the speaker's intention to produce a complete and structured text and are in fact the traces of language reflection:

*ona po-prezhnemu lysaya/i neponyatno rad on **ili net no-o/vidimo rad** tomu chto on//n-ne tak/poros* (S1, L1, w., 20, extr.);

no u nego v rukah uzhe/(laughing) *po-moyemu dve rascheski//**a mozhet byt' eto konechno i zerkalo/no vozmozhno eto takaya rascheska** znaesh' s pimpochkami takimi* (S4, L1, m., 21, extr.);

*pridya domoj on tut zhe rasstegnul sebe podtyazhki/snyal pidzhak i davaj na golovu znachit () (a) () vozle zerkala/**pered zerkalom** vylivat' etot eliksir* (S10, L1, w., 20, extr.).

Since self-repair is one of the characteristic phenomena of spontaneous speech, its usage can show the speakers' attitude toward their own speech and help to understand the difference of this attitude between representatives of different psychological groups.

The main psychological groups studied are introvert and extravert. Groups were formed based on EPI Questionnaire results.

An introvert is understood as a person prone to introspection, calm, restrained,

introjective and somewhat pessimistic, appreciating order, in need of a clear plan of action.

An extravert is understood as an impulsive, aggressive, optimistic and cheerful, outward-oriented personality, devoid of deep control of emotions.

Previous research in this area has shown that there is an explicit correlation between introversion/extraversion and the characteristics of spoken speech, especially when the research involves solving a difficult problem. Extraverts speak more, louder, with more repetitions, with fewer pauses and hesitations, they have a higher

[1] The elements of a repair are marked in all examples as follows: repaired text is underlined, the repair initiators are underlined and bold, repairing text is bold.

[2] The transcribing texts from the corpus "Balanced Annotated Text Library" include special symbols of discourse transcription: /—short physical pause; //—long physical pause; ()—short hesitation pause; (…)—long hesitation pause; (e-eh)—vocalized pause; - —prolongation of sounds; … – break, etc. For more on the transcription rules for the corpus "Balanced Annotated Text Library", see: [3].

[3] In the brackets after the example, we write the speaker's code, language group (native Russian speakers—"L1", Chinese Russian speakers—"L2"), gender, age, and psychological group (extravert, ambivert, or introvert).

speech rate and less formal language, while introverts have a wider vocabulary and speech is richer on average [4–6]. Extraverts also use more words to express positive emotions and demonstrate greater loyalty and quick compromise than introverts [7]. Researchers agree that the more difficult the task is and the higher the level of anxiety is expected during its completion, the easier it is to distinguish introverts from extroverts [8]. Most foreign researchers consider introversion/extraversion in the context of the "Big Five"—a personality model consisting of five common and relatively independent traits. Despite all attempts to find linguistic markers that fully reflect the psychological type of the speaker, the researchers did not consider self-repairs in this context.

The main purpose of the article is to reveal the correlation between the number of self-repairs in monologues and the psychological characteristics of the speaker. Moreover, the usage of a particular self-repair strategy by different psychological groups of speakers is also one of the study questions.

2 Self-Repairs and Their Classifications

A repair is one of the types of speech disfluencies marking the breaking point [9]. A self-repair is such a repair which "is produced by the speaker of the repairable" [10]. Syntactically, it is "any instance in which an emerging utterance is stopped in some way, and is then aborted, recast, or redone" [id: 80]. The main functions of self-repair are the indication to the interlocutor that one speech segment does not continue and how exactly the next segment should be understood regarding the previous segment [id.: 106].

The self-repair in particular and the process of speech monitoring for error correcting and speech editing in general includes three stages, according to Levelt [11]:

1. "monitoring of one's own speech and the interruption of the flow of speech when trouble is detected" [id.: 41];
2. marking the moment of interruption—the point at which the flow of speech is interrupted for 'editing' [12]—with the usage of editing terms (*uh*, *no*, *rather*, etc.);
3. making the repair proper.

C.L. Rieger calls these three stages the repairable, which is not necessarily presented in the text, "but can be inferred from the presence of repair initiation and the repairing segment" [13: 48], the repair initiation, which can consist of a cut-off, a filler, or a combination of these, and the repairing segment, which "repairs the trouble that the speaker has perceived" [ibid.].

B. A. Fox and R. Jasperson define these three text parts containing self-repair as repaired segment which is being repaired, repairable which refers to the repaired segment with the intention to alter it and can be missing (following Schegloff [14]), and repairing segment which accomplishes the repair [10].

There are two main types of self-repairs—error correction and self-editing. The distinction between error correction and repair as self-editing was firstly introduced by Schegloff et al. [15]. The correction is defined as one of several possible types of repairs which helps to replace an error by the correct word form. The repair is another type which serves to fix some disfluency in spontaneous speech.

Correction (online correction) and editing (retrospective correction) as two types of repairs can also differ by the time after which speakers consider the previous text inappropriate: correction arises immediately after the text which speakers want to change, it can break the ongoing discourse and its syntactic structure, and editing follows the completed fragment as a retrospective signal of change [9, 16].

Whereas "repair operates on and through syntax" [10: 82], the simple word change and phrase reformulation should be distinguished to analyze the main functions that self-repairs implement in spoken speech. In dialogues, such forms of self-repair can arise: word replacement, repairs on person reference, and repairs on next-speaker selection [15]. There is also a distinction between self-initiated repair when speakers repair the trouble in their own speech noticed by themselves and other-initiated repair when interlocutors point out that someone else's speech contains any kind of error or trouble [15, 17]. Since the material for this study is purely monological, all cases of repairs are self-initiated repairs which are implemented by the speakers themselves.

B.A. Fox and R. Jasperson classify self-repairs into seven different types:

Type A recycled word (word repetition);
Type B replaced word instead of another word or cut-off word;
Type C recycled phrase (phrase repetition);
Type D recycled phrase with one word replacement (part phrase repetition with a word replacement);
Type E recycled phrase with addition of new elements, which add background information;
Type F changing syntactic framework of recycled phrase;
Type G abandoning one structure and starting another [10].

All these operations are called "repeating or recycling, replacing or substituting, adding or inserting, and finally abandoning and restarting" [13: 50]. There are also different ways of restarting: instant repairs where the trouble word is traced and replaced, anticipatory repairs where the earlier word is traced and repeated, and fresh start where the restart with fresh material happens.

W.J.M. Levelt presented the typology of repairs which consisted of four groups: different, covert, appropriacy, and error repairs [11]. Different repairs (D) are repairs when the current phrase is replaced by a different one. Covert repairs (C) are such repairs when no morphemes are changed, added, or deleted, e.g., repetition of the same word.

Appropriacy repairs (A) include three categories: repairs that monitor the potential ambiguity in the context (AA-repairs), repairs that fix the use of appropriate level terminology (AL-repairs), and repairs that implement coherence with previously used expressions (AC-repairs). Appropriateness repairs usually are made for specification, not for correction itself. Error repairs cover lexical error repairs (EL-repairs), syntactic error repairs (ES-repair), and phonetic error repairs (EF-repairs).

S. Brédart includes in the group of appropriateness repair the new type—"repairs for good language"—which serves for social appropriateness of some part of the text which is being replaced [18].

Van Hest added to this taxonomy the group of conceptual error repair which is sometimes hard to differ from the lexical error repair [19].

J. Kormos divided all repairs into error repairs and inappropriate information repairs. She also introduced one category of repair—rephrasing repair [20]—which is highly important for the tasks of this research—since it is widely represented in spoken monologues.

Afterwards, Levelt's typology was slightly transformed [21, 22]. The new typology consists of two general categories: conceptualizer repairs (C-repairs), which implement pragmatic, semantic, and lexical changes, and formulator repairs (F-repairs) correcting grammatical and articulatory errors. Different, appropriacy, and lexical repairs were included into the category of C-repairs the main purpose of which is clarification using phrases reformulation or words editing (lexical, stylistic, or syntactic change). Pronunciation, morphology, and syntax repairs fall within the category of F-repairs which reflect the cases of pronunciation, morphological, or syntactic error correction. It basically supports the distinction between self-editing and error correction, but is hard to apply since there is no clear dividing line between lexical and appropriacy repairs.

Including the full words or phrases repetitions in the class of self-repairs is a common approach (see: [11, 13, 14, 23–25]), however, formally speaking, these repetitions reflect the speech spontaneity and are in fact one of the types of hesitative phenomena or speech disfluencies. In case there is no change in the following speech fragment, it seems that there is no verbalized self-repair, but only the buying-time repetition of what has been said, "marking time leads to overt 'search'" [14: 279]. We can only presume that the speakers wanted to make a change in their previous speech, but since there is no verbal "trace" of its kind, the self-repair remains unpronounced. In this article, we do not include full repetitions of cut-offs, words, or phrases in the group of self-repairs because of the reason for the absence of any change in word choice, word order, phrase structure, pronunciation, etc.

Repairs usually are introduced in the spoken discourse with the usage of different initiators. Non-lexical initiators can be cut-off, lengthening of sounds, and quasi-lexical fillers such as '*uh*' and '*um*' [14]. Lexical initiators observed are '*that is*', '*rather*', and '*I mean*' [26], '*well*', '*one moment*', '*whatchamacallit*', '*oh!*' [9], '*no*', '*not*', '*more correctly*', '*ah!*' [27]. They are also called "editing terms", for error repairs they are '*no*', '*sorry*', '*or*' [11].

Thus, based on findings of previous works, in this paper we present the new typology of self-repairs which focuses on three criteria: the type of repair (error correction or self-editing), the type of repair operation (word form change, word change, replacement of a word part with a word, word insertion, and phrase reformulation), and the type of repair initiation (with the usage of breaks, hesitation pauses, vocalized pauses, prolongations of sounds, sighs, word repetitions, part-word repetitions, and markers). In the next section all these types are described in detail.

3 Data and Processing Procedure

The data for the study are 24 monologues-descriptions in Russian. All the monologues belong to the corpus “Balanced Annotated Text Library”, which was created at the Philology Department of St. Petersburg State University. The corpus includes more than 700 monologues recorded from over 200 speakers that are native Russian speakers and speakers of Russian as a foreign language. The speakers were given the psychological H.J. Eysenck’s test to determine their extraversion/introversion level. All monologues were transcribed noting in the discourse annotation various types of pauses, including hesitation, paralinguistic elements (primarily laughter, sigh, cough), breaks, prolongation of sounds, and other signs of spontaneity [2, 28].

The material for the research consists of two groups of texts: 12 monologues of native Russian speakers that were students of Saint Petersburg State University Philology Department and 12 monologues of foreign Russian speakers that were Chinese students that studied at Saint Petersburg State University Philology Department. The total number of tokens in monologues produced by native speakers is 2456, the total number of tokens in monologues produced by Chinese speakers is 1767. The sample includes 8 extraverts, 13 introverts, and 3 ambiverts, but the word material of extraverts and introverts is balanced. Extraverts’ monologues in total include 1868 tokens, introverts’ monologues include 1814 tokens, ambiverts’ monologues—541 tokens. The sample is gender-balanced; all the speakers represent the young age group (from 18 to 27 years old). Since the source of the material is the corpus “Balanced Annotated Text Library”, which represents isomorphic texts recorded from different groups of native Russian speakers and foreigners speaking Russian as their second language, we suppose that it provides an opportunity to investigate these monologues seamlessly from the psycholinguistics perspective. The type of monologues remains the same for both groups: the stimuli for monologues production were the comic strips by H. Bidstrup “Hair loss treatment” [29]. Comparison of such texts with each other, taking into account the psychological characteristics of the speakers, can give interesting results, the value of which significantly increases if speakers who are known to have different linguistic experiences are involved in the study, that is, who speak the language (in this case, Russian) as their native language and as the second one. The comparative analysis of self-repairs in the speech of native and foreign speakers, as well as the investigation of any other units of spontaneous speech production, allows identifying, on the one hand, common, universal features of spoken discourse and, on the other hand, its peculiar characteristics.

In the monologues-descriptions, the cases of self-repair were annotated marking two types of self-repairs: error correction (morphological, syntactic, and pronunciation errors) and self-editing:

muzhchina kupil () elis… ***eliksir*** *dlya vOlosy* (S10, L2, m., 25, extr.)—pronunciation error correction;

krasivaya zhenshchina/y-n () posove… posovetovala () yemu chto () (cough) *() odnu-u* ***() (y-n) () lekarst… o… o… odno*** *lekarstvo* (S1, L2, w., 23, extr.)—morphological error correction;

on smotrelsya na-a () ***v zerkalo/****(y-n) () volosy yeshchyo dlinnee* (S4, L2, w., 24, intr.)—syntactic error correction;

ochen' lysyj muzhchina//pokupaet v magazine kakoj-to/kakoj-to eliksir//kak ponyat… ***kak () vidno*** *iz nazvaniya dlya volos* (S2, L1, m., 20, extr.)—self-editing;

potom/on zameti… (y) ***on prosnula/****zametil chto/u-u nego yest' volosy* (S10, L2, m., 25, extr.)—self-editing.

Besides, lexical and non-lexical means of self-repairs were also tagged: repair initiators and supporters—breaks, hesitation pauses, vocalized pauses, prolongations of sounds, sighs, word repetitions, word repetition after word break, conjunctions (*'ili'* (*'or'*)), and discourse markers (*'v obshchem'* (*'in general'*), *'tochneye', 'verneye'* (*'more correctly'*)),—and repair operation types: word form change, word change, replacement of a word part with a word (in case they are not identical), word insertion, and phrase reformulation. This classification of self-repairs basically refers to C.L. Rieger's types of repair operation—replacing, inserting, and abandoning [13: 50]—except for repeating:

cherez dve nedeli/u neyo ***u nego*** *dejstvitel'no poluchilis' o… ochen' gustye () shikarnye () volosy* (S7, L2, w., 25, intr.)—word form change with the preposition repetition;

(a-a) u menya volos yeshchyo ***(…) (e-e) poka*** *ne ostalos' na golove* (S12, L2, w., 26, amb.)—word change after the silent and vocalized pauses;

a vot v etom moment ya vst… ***ya-a () prosnulsya*** * *() zasnulsya* ** *iz-za uzhasov* (S12, L2, w., 26, amb.)—* replacement of a part of one word with another word implemented by the cut-off, repetition of the pronoun with the prolongation, and hesitation pause, ** word form change after the silent pause;

*za noch' proizojti i lozhitsya spat'/i che…/****trogaet*** *svoyu golovu/chto n-ne n-nachali li rasti u nego volosy* (S1, L1, w., 20, extr.)—replacement of a part of one word with another word, means of repair are the cut-off and the pause;

poslednyaya kartinka v (e) ***pravaya nizhnyaya*** * *v poslednem nizhnem* (sigh) *ryadu/govorit nam o tom* ***ili vernee ne govorit/a zastavlyaet*** ** *nas usomnit'sya v tom a bylo li eto pravdoj/ili eto byl vsyo-taki son (e) smeshnogo cheloveka kotoryj () zahotel imet' (e-e) prekrasnuyu shevelyuru* (S9, L1, m., 20, intr.)—* words insertion introduced with the vocalized pause (*eh*), ** word change introduced with the marker *'ili vernee'* (*'or more correctly'*);

eto-o () ochen' (…) (e-n) (…) g… ***() (e-y) () (e-y) () yemu ochen' grustno*** (S1, L2, w., 23, extr.)—phrase reformulation using the cut-off and the chain of hesitation pauses.

Table 1 Self-repairs in spoken discourse of extraverts, introverts, and ambiverts

Psychological group	Number and %
Extraverts	34
Introverts	50
Ambiverts	15

Moreover, different types of errors (pronunciation, lexical, morphological, syntactic, and stylistic) were also revealed in monologues to understand the frequency of actual error correction.

The data annotation involved as well the meta-data notion (information about gender, age, psychological characteristics, and, for foreign speakers, information on their level of language proficiency). One way or another, the distribution of informants by psychotype, according to the EPI test results, turned out to be insufficiently uniform in order to talk about psycholinguistic balancing. The prevailing number of introverts, according to the test results, is explained, first of all, by the fact that psychological testing was carried out after the recording and could not be a reason for refusing or inviting a particular speaker to participate in the experiment.

4 Correlation Analysis

The present study was based on all the criteria necessary for the successful application of methods of quantitative linguistics in practice and in application to linguistic theory, despite the fact that they are usually correlated exclusively with the practical field. This approach makes it possible to significantly expand and modify the scientific picture of the entire language system and the possibilities of its functioning [30].

Statistics and quantitative methods allow us to effectively analyze linguistic data in order to draw conclusions based on mathematical values: intuitive comparisons are not always reliable and often are not able to highlight significant features, to show whether the differences existing in the data can be accidental. All further analysis was carried out with the help of computer methods, using the Python programming language, in particular, the libraries for analyzing large data arrays NumPy and Pandas.[4] To build correlation matrices, were also used Matplotlib and Seaborn.[5]

The total number of self-repairs in the analyzed data is 99 cases used by 20 speakers (4 speakers did not use any self-repairs in their monologues). The total number of self-repairs in spoken speech produced by extraverts, introverts, and ambiverts is shown in Table 1.

[4] All software used in this study is free and distributed as open source: https://pandas.pydata.org/, https://numpy.org/.

[5] Documentation for these two libraries is available at: https://matplotlib.org/ and https://seaborn.pydata.org/.

In the data investigated, half of all self-repairs belong to introverts’ speech. However, it is important to analyze which type of self-repair is more often used by representatives of different psychological groups. Table 2 shows the number of error corrections and self-editing repairs among speakers from each group.

In the analyzed data, there were 286 errors (13 were in native Russian speakers’ monologues and 273 were in foreign Russian speakers’ monologues) and 38 actual error corrections. So, only 13% of errors were corrected by the speakers. It can be seen from Table 2 that self-editing is a more frequent type of self-repair (62% vs. 38%), which is used almost evenly often both by extraverts (42% of all cases of self-editing) and introverts (47% of all cases of self-editing). Error correction is especially often used by introverts (58% of all error correction cases). Thus, we can assume that self-editing is a more common self-repair type used by different speakers, and error correction is more important for introverts, which tend to notice and repair actual errors in their speech:

*on dovol’nyj idyot domoj//ves’/flakonchik//**kogda nu/pridya domoj on-n/vylivaet ves’ fl…/flakonchik** sebe na golovu* (S2, L1, m., 20, extr.)—self-editing;

*vdrug on podoshyol k ze… **on vzyal zerkalo** i smotrel na sebe* (S8, L2, m., 23, amb.)—self-editing;

*yemu prihoditsya () (y-n) () (n) narezat’ **() (y-n) rezat’** volosy/no (n) volosy yeshchyo () (n) rastyo… () rastyot* (S6, L2, w., 24, intr.)—error correction.

For the further analysis, spoken monologues of representatives of three psychological groups—extraverts, introverts, and ambiverts—were compared by the main feature—repair operation types revealed in the material. The total distribution of different repair operations in spoken speech is shown in Table 3.

Table 2 Error correction and self-editing in spoken discourse of extraverts, introverts, and ambiverts

Psychological group	Error correction		Self-editing	
	Number	Percentage of all error corrections	Number	Percentage of all self-editing
Extraverts	8	21	26	42
Introverts	22	58	28	47
Ambiverts	8	21	7	11

Table 3 Self-repair operations in spoken discourse

Repair operation	Number and %
C (word change)	16
F (word form change)	27
I (word insertion)	16
N (word part replacement)	31
R (phrase reformulation)	9

The most frequently used types of self-repairs among all the speakers are replacement of a word part with a word (N, 31%) and word form change (F, 27%). The analysis shows that speakers tend to cut off the word that was chosen incorrectly, as they might think, as fast as possible, not continuing to pronounce it completely, cf. Main Interruption Rule: "Stop the flow of speech immediately upon detecting the occasion of repair" [11: 56]. The full-pronounced mischosen words are twice less frequent than replacement of cut-offs (C, 16%), as well as insertion of a word in an ongoing phrase using the repair (I, 16%).

To reveal the possible correlation between the psychological characteristics of the speaker and the self-repair operation type used, the number of such cases in texts produced by representatives of each psychotype was counted. The comparison of the number of repair operations used by representatives of three groups is shown in Table 4.

Introverts in general use more self-repairs than representatives of other psychological groups. However, the most frequent repair operations among introverts are replacement of a word part with a word (N, 19%) and word form change (F, 14%), which lead by a large margin. By contrast, extraverts do not have a big difference between repair operations number, but the most frequent type turned out to be word change (C, 9%). Extraverts in their monologues more often complete the mischosen word, finish to pronounce it without a break, although introverts prefer short word cut-off and its replacement with the right word. Phrase reformulation is the least common type of repair operations in speech of all psychological groups. Thus, there

Table 4 Self-repair operations in spoken discourse of extraverts, introverts, and ambiverts

Psychotype	Repair operation	Number and %
Extraverts	C (word change)	9
	F (word form change)	8
	I (word insertion)	7
	N (word part replacement)	6
	R (phrase reformulation)	4
Introverts	C (word change)	5
	F (word form change)	14
	I (word insertion)	8
	N (word part replacement)	19
	R (phrase reformulation)	4
Ambiverts	C (word change)	2
	F (word form change)	5
	I (word insertion)	1
	N (word part replacement)	6
	R (phrase reformulation)	1

is less need to repair the syntactic surface; however, the right word choice, according to the data, is more important for all the speakers in terms of self-repair usage.

Self-repairs, for the most part—97% of all examples,—are introduced in speech flow by the repair initiators, such as pauses, breaks, prolongations, etc., and accompanied by the repair supporters, e.g., word or word part repetitions. To reveal the total number of different repair initiators and supporters in spoken data, each repair mean in a chain of several repair implementing means was counted separately. The total number of different repair means is presented in Table 5.

In total, 206 different repair means were used in the analyzed material. The most common among all the speakers turned out to be silent pauses (P, 35.9%), breaks (B, 25.7%), and vocalized pauses (V, 12.6%).

For the comparison of the number of repair initiators and supporters used by extraverts, introverts, and ambiverts, each repair mean was counted separately as well. Table 6 shows the number of different repair means used by speakers from three psychological groups.

Silent (P, 18.45%) and vocalized (V, 9.22%) pauses, as well as breaks (B, 14.56%) are widely used by introverts as repair initiators; extraverts more often use discourse markers (D, 4.37%) than introverts and ambiverts and strictly prefer silent pauses (P, 12.14%) and not vocalized pauses (V, 1.94%). Ambiverts tend to use breaks (B, 4.85%) and silent pauses (P, 5.34%) before self-repairs as repair initiators. The number of word repetitions (E), part word repetitions (A), conjunctions (J), and prolongations (L) is approximately equal in speech of extraverts and introverts and usage of these means does not correlate with the speaker's psychotype.

The repair initiators tend to be used in monologues in a chain, e.g., break, repetition, prolongation of a sound, pause, and repair:

nu vy znaete/ya zhe-e () syto ***… ya zhe-e () lysyj*** *() ve-e () takoj rascvetogo vozrast* (S12, L2, w., 26, amb.).

Table 5 Self-repair means in spoken discourse

Repair mean	Number	Percentage to total
A (word repetition after word break)	6	2.9
B (break)	53	25.7
D (discourse marker)	10	4.9
E (word repetition)	24	11.7
H (sigh)	1	0.5
J (conjunction)	5	2.4
L (prolongation)	7	3.4
P (silent pause)	74	35.9
V (vocalized pause)	26	12.6
TOTAL	206	100

Table 6 Self-repair means in spoken discourse of extraverts, introverts, and ambiverts

Psychotype	Repair mean	Number	Percentage to total
Extraverts	A (word repetition after word break)	2	0.97
	B (break)	13	6.31
	D (discourse marker)	9	4.37
	E (word repetition)	10	4.85
	J (conjunction)	2	0.97
	L (prolongation)	3	1.46
	P (silent pause)	25	12.14
	V (vocalized pause)	4	1.94
Introverts	A (word repetition after word break)	3	1.46
	B (break)	30	14.56
	D (discourse marker)	1	0.49
	E (word repetition)	12	5.83
	J (conjunction)	3	1.46
	L (prolongation)	2	0.97
	P (silent pause)	38	18.45
	V (vocalized pause)	19	9.22
Ambiverts	A (word repetition after word break)	1	0.49
	B (break)	10	4.85
	E (word repetition)	2	0.97
	H (sigh)	1	0.49
	L (prolongation)	2	0.97
	P (silent pause)	11	5.34
	V (vocalized pause)	3	1.46

In an attempt to comprehend the interactions of the self-repairs (for example, what repair means are used in one chain) in the speech of informants belonging to the same psychological type, a correlation analysis was also carried out with the calculation of the linear correlation coefficient r-Pearson. Figure 1 presents the correlation matrix as the result of analysis of co-occurrence of different repair means for introverts. Matrix as Fig. 2 presents the same for extraverts. R-Pearson coefficient was used as a measure of how close the different self-repair types are to a line of best fit. The Pearson correlation coefficient also tells whether the slope of the line of best fit is negative or positive.

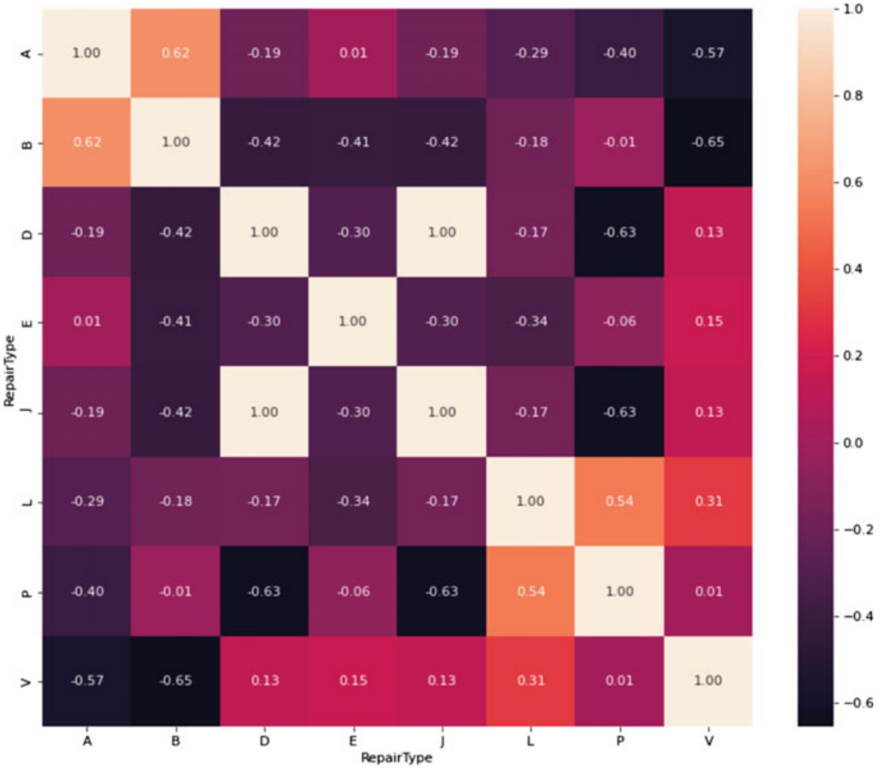


Fig. 1 Correlation between different self-repair means (introverts)

Since the tasks of this study were not intended to obtain such estimates, we do not dwell on the analysis of these results in detail. In the perspective of research in this area, it is proposed to consider the following range of hypotheses formulated on the basis of matrices:

1. in self-repair chains introverts often use discourse markers along with conjunctions;
2. in self-repair chains introverts never use a break and a vocalized pause simultaneously;
3. in self-repair chains extraverts never use a discourse marker along with a silent pause and a break and a conjunction simultaneously.

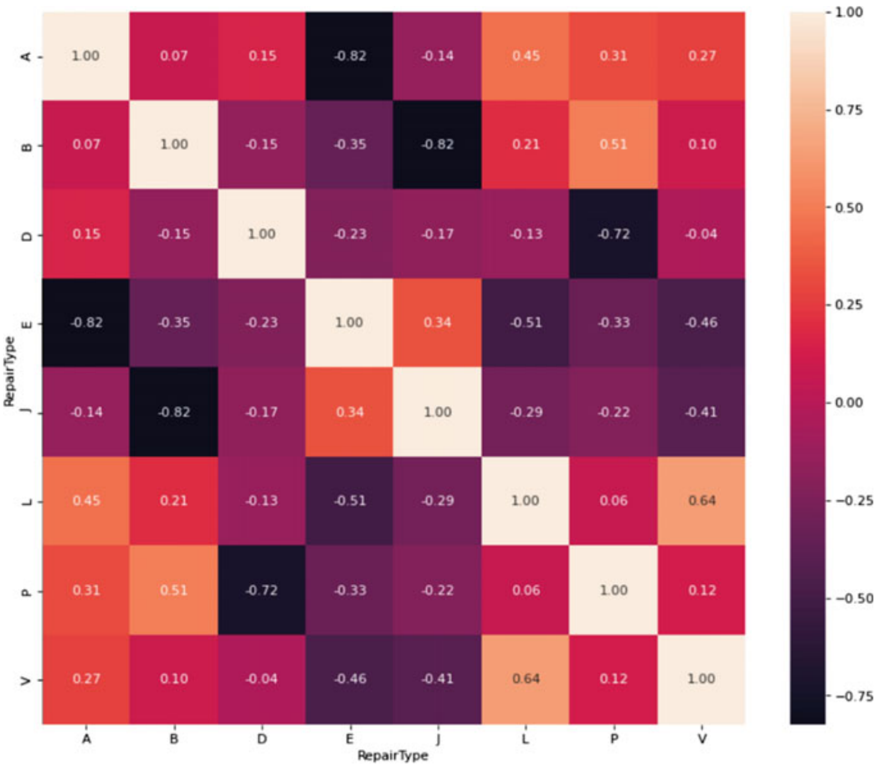


Fig. 2 Correlation between different self-repair means (extraverts)

5 Conclusion

Self-repairs play an important role in spontaneous speech production of all psychological groups of speakers. The usage of a particular self-repair strategy by representatives of a certain psychological group can show how the speakers' psychotype can affect their speech production. Since self-repair is one of the characteristic phenomena of spontaneous speech, its usage can show the speakers' attitude toward their own speech and help to understand the difference in this attitude between representatives of different psychological groups.

For this research, the new typology of self-repairs was designed based on three criteria: the type of repair (error correction or self-editing), the type of repair operation (word form change, word change, replacement of a word part with a word, word insertion, and phrase reformulation), and the type of repair initiation (with the usage of breaks, hesitation pauses, vocalized pauses, prolongations of sounds, sighs, word repetitions, part-word repetitions, and discourse markers). All the cases of self-repairs were annotated following this typology.

Thus, as a result, it turned out that there is a difference in self-repairs' usage among extraverts, introverts, and ambiverts. Half of all self-repairs belong to introverts' speech, where error correction is especially frequent. The most often used repair operations among introverts are replacement of a word part with a word and word form change. Extraverts' most frequent type of repair operations is word change. We can conclude that extraverts in their speech more often finish the mischosen word, however, introverts cut the wrong word off and replace it with the right one. Almost all self-repairs are introduced in speech flow by the repair initiators, such as pauses, breaks, prolongations, etc. Introverts initiate their repairs with breaks, silent and vocalized pauses, extraverts only with silent pauses, but more often use discourse markers. Ambiverts tend to use breaks and silent pauses before self-repairs as the most frequent repair means. Therefore, in the case of self-repair, we can see the correlation between the extraversion/introversion level of speakers and the features of their speech production. Further investigation may involve the expansion of analyzed data, the exploration of self-repair initiation chains among different speakers, and, thus, better understanding of the self-repair process and results.

Acknowledgements The presented research was partially supported by Saint Petersburg State University, project #92562973 "Modeling of Russian Megalopolis Citizens' Communicative Behavior in Social, Speech and Pragmatic Aspects Using Artificial Intelligence Methods".

References

1. Shriberg, E. E. (1994). *Preliminaries to a theory of speech disfluencies*. Ph.D. dissertation. University of California at Berkeley.
2. Sound corpus as a material for analysis of Russian speech: Collective monograph. Part 1. Reading. Retelling. Description. In N.V. Bogdanova-Beglarian (Ed.). St. Petersburg State University, Philology department, St. Petersburg (2013). (in Russian).
3. Zaides, K. D. (2019). On unification of annotation of the "Balanced Annotated Text Library" corpus. In *Proceedings of International Conference "Corpus Linguistics–2019* (pp. 332–339). St. Petersburg University Publ., St. Petersburg. (in Russian).
4. Scherer, K. R. (1979). Personality markers in speech. In K. R. Scherer & H. Giles (Eds.), *Social markers in speech* (pp. 147–209). Cambridge University Press.
5. Furnham, A. (1990). Language and personality. In H. Giles & W. P. Robinson (Eds.), *Handbook of language and social psychology* (pp. 73–95). John Wiley & Sons.
6. Gill, A., & Oberlander, J. (2002). Taking care of the linguistic features of extraversion. In *Proceedings of the 24th Annual Conference of the Cognitive Science Society* (pp. 363–368). Cognitive Science Society.
7. Pennebaker, J. W., & King, L. (1999). Linguistic styles: Language use as an individual difference. *Journal of Personality and Social Psychology, 77*, 1296–1312.
8. Dewaele, J. M., & Furnham, A. (2000). Personality and speech production: A plot study of second language learners. *Personality and Individual Differences, No, 28*, 355–365.
9. Kibrik, A. A., & Podlesskaya, V. I.: Speaker's self-repairs and other types of speech failures as an object of annotation in spoken speech corpora. In *Scientific and technical information. Series 2: Information processes and systems* (vol. 2, pp. 2–23). All-Russian Institute for Scientific and Technical Information Publ. (in Russian).

10. Fox, B. A., & Jasperson, R. (1995). A syntactic exploration of repair in English conversation. In P.W. Davis (Ed.), *Alternative linguistics: descriptive and theoretical modes* (pp. 77–134).
11. Levelt, W. J. M. (1983). Monitoring and self-repair in speech. *Cognition, 14*(1), 41–104.
12. Hockett, C. F. (1967). Where the tongue slips, there slip 1. In *To Honor Roman Jakobson* (vol. II, pp. 910–936). Mouton.
13. Rieger, C. L. (2003). Repetitions as self-repair strategies in English and German conversations. *Journal of Pragmatics, 35*(1), 47–69.
14. Schegloff, E. A. (1979). The relevance of repair to syntax-for-conversation. In T. Givón (Ed.), *Discourse and syntax* (pp. 261–286). Academic Press.
15. Schegloff, E. A., Jefferson, G., & Sachs, H. (1977). The preference for self-correction in the organization of repair in conversation. *Language, 53*, 361–382.
16. Podlesskaya, V. I. (2015). A corpus-based study of self-repairs in Russian spoken monologues. *Russian Linguistics, 39*, 63–79.
17. Schegloff, E. A. (2007). *Sequence organization in interaction: A primer in conversation analysis* (Vol. 1). Cambridge University Press.
18. Brédart, S. (1991). Word interruption in self-repairing. *Journal of Psycholinguistic Research, 20*, 123–138.
19. Van Hest, H. E. (1996). *Self-repair in L1 and L2 production: An overview*. Tilburg University Press.
20. Kormos, J. (1998). A new psycholinguistic taxonomy of self-repairs in L2: A qualitative analysis with retrospection. In *Even Yearbook, ELITE SEAS Working Papers in Linguistics* (vol. 3, pp. 43–68).
21. Simard, D., Fortier, V., & Zuniga, M. (2011). Attention et production d'autoreformulations autoamorcées en français langue seconde, quelle relation? *Journal of French Language Stu-dies, 21*(3), 417–436.
22. Zuniga, M., & Simard, D. (2019). Factors influencing L2 self-repair behavior: The role of L2 proficiency, attentional control and L1 self-repair behavior. *Journal of Psycholinguistic Research, 1*, 43–59.
23. Fox, B. A., Hayashi, M., & Jasperson, R. (1996). Resources and repair: A cross-linguistic study of syntax and repair. In E. Ochs, E. A. Schegloff, & S. A. Thompson (Eds.), *Interaction and grammar* (pp. 185–237). Cambridge University Press.
24. Fox, B. A., Maschler, Y., & Uhmann, S. (2010). A cross-linguistic study of self-repair: Evidence from English, German, and Hebrew. *Journal of Pragmatics, 42*, 2487–2505.
25. Bada, E. (2010). Repetitions as vocalized fillers and self-repairs in English and French interlanguages. *Journal of Pragmatics, 42*(6), 1680–1688.
26. DuBois, J. W. (1974). Syntax in mid-sentence. In *Berkeley studies in syntax and semantics* (vol. 1, pp. III.1–III.25). University of California, Institute of Human Learning and Department of Linguistics.
27. Zaides, K. D. (2016). Meta-communicative insertions in Russian oral spontaneous speech of native speakers and foreigners. *Communication Studies, 3*(9), 19–35. (in Russian).
28. Bogdanova-Beglarian, N. V., Blinova, O. V., Sherstinova, T. Yu., & Zaides, K. D. (2019). Corpus «Balanced Annotated Text Library»: Analysis of Russian monological speech. In *Proceedings of the V.V. Vinogradov Russian Language Institute* (pp. 111–126). RLI RAS. (in Russian).
29. Hair loss treatment. https://herlufbidstrup.com/comics/Hair_loss_treatment. Retrieved 14 May 2022.
30. Baranov, A. N. (2001). *Introduction to applied linguistics*. Yeditorial URSS. (in Russian).

Den’ Den’skoj: A Lexicographic Portrait of a Russian Microsyntactic Unit

Alexandra Chaga

Abstract Microsyntactic units (MSU), such as syntactic idioms and non-standard syntactic constructions, present a significant yet not sufficiently investigated area of language phenomena. In this paper we focus on *den’-den’skoj* (≈‘day-to-day’)—an MSU containing repeated elements. The study of *den’-den’skoj* construction illustrates a microsyntactic approach, which involves identification and full description of specific MSUs, as well as the development of two linguistic resources: a Microsyntactic Dictionary, and a microsyntactically marked-up corpus, where the MSUs are indicated and assigned particular meanings. A full lexicographic portrait of an MSU includes a lexicographic definition, a structured description specifying all morphological and syntactic parameters, valence properties, combinatorial possibilities and semantic features. Microsyntactic units with repeated elements present an outstanding kind of MSU’s. On the one hand, they involve duplication of various types, which is often considered a lexical error. On the other hand, as duplicated expressions become fixed, some of them are no longer regarded as incorrect. Over time, they become naturally usable in various contexts, sometimes generating new idiomatic expressions.

Keywords Russian · Microsyntax · Construction with repeated elements

1 Introduction

This paper reports new results of research into Russian microsyntactic units (MSU). The research has been carried out for two decades at the Laboratory of Computational Linguistics of the Harkevič Institute for Information Transmission Problems, of the Russian Academy of Sciences.

The very term “microsyntax” emerged from the idea that natural language syntax includes a subset of phenomena that stand out from the general realm of syntax.

A. Chaga (✉)
Institute for Information Transmission Problems, Russian Academy of Sciences, 103051 Moscow, Russia
e-mail: chaga@iitp.ru

P. Eismont et al. (eds.), *Literature, Language and Computing*,
https://doi.org/10.1007/978-981-99-3604-5_2

This subset involves specific lexical units or narrow classes of such units and reflects peripheral and clearly language-specific meanings, such as Russian *dat' otmašku* 'give the go-ahead', *v obŝem* 'all in all', *vo vsjakom slučae* 'in any case'. The elements that make up the set of these phenomena are opposed to the "large" basic syntax of the language. Historically, this name goes back to the English term **minor type sentences**, which gained some spread during the 1970s–1980s. For some time, this set of phenomena was referred to as "minor syntax", but the term "microsyntax" was considered more appropriate for its author, Leonid Iomdin. Besides the idea of a minor scale, it alludes to much more precise tools and methods of research to be used, just as microsurgery requires much more precise instruments than general surgery does.

Sure enough, the study of these phenomena is much older than the term "microsyntax". Thus, analyzing some representative types of Russian language constructions, Švedova [1] and Šmelëv [2] named them "phraseoschemata". Melčuk [3, 4], used the term "syntactic phrasemes", which echoed the concept of "syntactic idioms" [5]. For some of these constructions, Apresjan and Iomdin introduced the term "syntactic agglomerates" [6, 7]. Later on, Iomdin proposed the term "microsyntax" to describe a wide range of syntactic-semantic phenomena and has actively worked in this field [8, 9].

The area of microsyntax embraces a large variety of units, each of which has its own unique structure and distinctive characteristics that reflect different lexical meanings or display hardly predictable syntactic properties [10].

A microsyntactic approach presupposes identification and full description of specific MSUs, as well as the development of high-quality linguistic resources that are integrally related to each other: a Microsyntactic Dictionary, which is largely based on the ideas of the Active Dictionary by Apresjan and his colleagues [11], as well as a microsyntactically marked-up corpus, where the microunits are determined and assigned particular meanings. In the nearest future, a microsyntactically annotated corpus, SynTagRus, containing more than 36,000 microunit entries, will become available through the website www.ruscorpora.ru, the Russian National Corpus [12].

According to the degree of lexicalization, we distinguish between two main groups of idiomatic linguistic units: (1) weakly lexicalized or lexically unaffected non-standard syntactic constructions (X X-u rozn' $\approx$ 'one X is different from another X': *čelovek čeloveku rozn'* 'people are different'*)* and (2) lexically restricted syntactic idioms (*net-net da i* 'occasionally, from time to time', as in *devočka tut est', sirotka: net-net da i navedaetsja* (Ivan Turgenev) 'There's a little girl here, an orphan; now and then she comes to see me' (RNC, Russian-English parallel corpus).

It is worth saying that there are no clear boundaries between the two types, and a considerable number of microsyntactic units are somewhere in the middle. Also, many constructions are equally related to microsyntax and to the area of classical phraseology, especially to the grammatical phraseology in its broader sense, including not only morphological, but also syntactic and lexical phenomena of natural language [13]. The main criteria used to identify an MSU is its non-standard syntactical behavior and, frequently, irregularity of the ways of expressing grammatical meanings.

At the moment, the list of microsyntactic units presented in SynTagRus consists of a little more than 3000 different elements, and a considerable part of these are MSUs with repeated lexical elements such as *v konce koncov* 'in the end'*, vremja ot vremeni* 'from time to time'*, so dnja na den'* 'any day'*, delo est' delo* 'business is business', *hudožnik na to i hudožnik* 'this is what an artist is for' etc.

Studies devoted to a variety of Russian repetitive expressions have been regularly published lately; however, the material has not received a systematic presentation yet due to its peripheral status in the grammar. The microsyntactic approach includes lexicographical description (or a lexicographical portrait—term coined by Apresjan [14]) for each microunit. This means a structured description with specification of all morphological and syntactical parameters, as well as its lexicographical definition, its valence properties, combinatorial possibilities and semantic features.

Pleonasms can be observed in all microsyntactic units with repeated elements. On the one hand, there is a duplication of some component of meaning and a repeated expression of the same meaning within one text segment, which is often considered a lexical error (*maslo masljanoe* 'oily oil / buttery butter'). On the other hand, some expressions of this kind become fixed and as such are no longer regarded as incorrect. What is more, they are sustainably used over time, finding their natural place in various contexts. Some of them encourage the emergence of new expressions: *čudo čudnoe* 'wonderful wonder' or *užas užasnyj* 'terrible terror'.

2 Syntactic Idiom Den'-Den'skoj

Den'-den'skoj is formally constructed by the scheme X X-ovyj, like *čudo čudnoe / čudo čudesnoe* 'wonderful wonder'), *muka mučeničeskaja* 'anguished anguish'), *dali dal'nie* 'far far-aways' etc. In the scheme, the variable X represents a noun, and X-ovyj—a same root adjective agreed with X.

Some examples from the Russian National Corpus (RNC) [12] illustrate the construction:

(1) *Čudnoe poistine mesto—obryvistyj mysok nad vodoj, otkuda vidno tak široko i mnogo, čto* ***den' by den'skoj*** *sidel i gljadel by.* 'It's a truly wonderful place—a steep cape above the water, from where you can see so widely and so much that you would sit and look like that all day long.'

(2) *A den'gi gde vzjat', esli ne u materi? Vot i topaju* ***den'-den'skoj****,* ***noč-nočen'skuju****. Nogi opuhat' stali.* 'And where could I get the money if not from my mother? So, I would stomp around all day long and all night long. My feet became swollen.'

(3) *Bez umolku* ***den'-den'skoj*** *šumel les, a pridët noč, zagorjatsja zvëzdy, i v zvëzdah, kak car', gudit les grozno, volnuetsja.* 'All day long the forest roared, and when night would come, the stars would light up, and in the stars, like a king, the worried forest hummed menacingly.'

2.1 Morphological Characteristics

The contexts where this construction occurs reveal that it is used exclusively in the accusative case, singular number, which is quite natural when denoting duration (discussed below). Some rare exceptions display a play of words or a stylistic device:

(4) *Tak i korotaju* ***dni-den'skie***. 'And so I pass the days.' (ARC)

or

(5) *A za barhatom štor bušuet* ***den'-den'skoj***. 'And behind the velvet curtains rages the bright day' (RNC)

It seems that the constraints imposed on the syntactic position are more rigid than those imposed on the number. The phrase looks less acceptable when *den'-den'skoj* is functioning as a subject than when it is a complement of duration:

?? ***Den'-den'skoj*** *prošël v hlopotah.* 'The whole day was full of chores.'

vs. ? ***Dni-den'skie*** *ona provodit v hlopotah.* 'She spends all day long in chores'.

Unlike microunits type of *čudo čudnoe, den'-den'skoj* is usually hyphenated. In Araneum Russicum corpus (ARC) (http://ucts.uniba.sk/aranea_about/_russicum.html), the hyphenated form occurs 9 times more frequently than the non-hyphenated form. This indicates a higher internal cohesion of the components, which, however, does not prevent the construction from being split:

Den'-to den'skoj *ona u plity stoit* 'All day long she is cooking'*;* or ***Den' že den'skoj*** *ona valjaetsja na divane.* 'All day long she is lying on the sofa'.

The element *den'skoj* is an adjective, which is shown by its morphological features. As a result of the fixed word order (**den'skoj den'*) and a special adjectival form, which is never used apart from the noun *den*', the construction's degree of semantic idiomaticity is rather high. Note that *X X-ovyj* constructions, which are close in form, allow changing the word order and inserting not only particles but also verbs:

žutkaja žut', 'terrible terror'*, čudo slučilos' čudnoe* 'a wonderful wonder happened'.

The microunit *den'-den'skoj* has a special suffix -*sk*, which, on the one hand, removes the negative effect of tautology, and on the other hand highlights some, though not all, the components of meaning of the motivating noun *day*. Thus, expressions with the basic suffix -*n* like**den'-dnevnoj* or * *dnevnoj den'* hardly occur in texts. The reason for that is that the conventional form of the adjective duplicates the motivating noun's sense, whereas the element *den'skoj* highlights the component of duration, limited only by the daylight hours, and does not correlate with the day in the meaning '24 h'.

Basically, the grammar does not prohibit the construction having other inflectional forms than genitive singular, eg.: *dni-den'skie* (nominative case, plural number), *dnej-den'skih* (genitive case, plural number), *dnjam-denskim* (dative case, plural number) etc. All these forms seem to be acceptable. The defective paradigm is caused by the syntactic restrictions. At the same time, cognate constructions built on the *X X-ovyj* pattern, are freely used in various cases:

(6) *«Nazovite* ***čudom čudnym*** (DAT)*, nazovite, kak hotite»,—dobavila ona, vspomniv o svoej fraze, skazannoj v odnom iz prošlyh èfirov.* "Call it a wonderful wonder, call it what you want," she added, recalling her phrase from one of the previous broadcasts.' (RNC).

(7) *Zato kak otletit v* ***dali dal'nie*** (ACC)*, v dumy tvorčeskie, to i ne vspomnit ni o kakom takom slučae, i ljudi radujutsja, gljadja na èkran ili na scenu: on li èto?* 'But as soon as he flies off into distant places, into his creative thoughts, he wouldn't remember such an incident, and people would rejoice when they look at the screen or at the stage wondering if it is him.' (RNC).

2.2 Semantic Properties

The meaning of the *X X-ovyj* syntactic scheme can be described as intensification of the basic meaning of the noun, which is emphasized by the dependent same root adjective. Along with the intensification and emphasis there is often an element of slightly positive assessment or empathy with the protagonist of the situation. Even constructions like *žut' žutkaja* 'horrible horror' or *užas užasnyj* 'terrible terror' mainly refer to emotional intensity rather than to an enhanced state of horror or terror. Moreover, in colloquial speech based on this scheme, constructions with more likely positive evaluations are built on the fly: e.g. *krasota krasivaja* 'beautiful beauty' and *prelest' plelestnaja* 'charming charm'. Rather, overtly negative connotation is evoked by constructions like *X X-om,* with the second instance of X appearing in the instrumental case: *drjan' drjan'ju* 'trashed trash', *durak durakom* 'foolish fool'.

Compared to *X X-ovyj*, the semantics of *den'-den'skoj* is more complex. It is close to the expressions *celyj den', den'(dni) naprolët* 'all day(s) long' and *každyj den'* 'from day to day'. *Den'-den'skoj* describes some long-lasting events or states that occur during the daylight time, sometimes with an emphatic meaning of regular repetition and routine. More often it implies actions unlikely to be completed in one day. The occurrences of this phraseme demonstrate that *den'-den'skoj* refers specifically to the events of daylight hours, and does not refer to nights or evenings. There are quite a few examples in the Russian National Corpus [12] (https://ruscorpora.ru), where events characterized by the adverbial *den'-den'skoj* are contrasted with events of the night. Sometimes this gives rise to author's expressions like *noč-nočen'skaja* 'all night long', but more often expressions like 'all the night', 'at night' etc. are used in this context. However, there are frequent examples with no opposition to the night, yet containing explicit limitations of the day:

(8) *Predstavim sebe, čto značit prorabotat' v takoj atmosphere* ***den'-den'skoj****, dopozdna.* 'Imagine what how does it feel to work in such an atmosphere all day long, late into the night.' (RNC).

Beyond that, there is a component of exhaustive completeness in the microunit *den'-den'skoj*. While describing an event, lasting continuously from morning to evening, the speaker introduces a specific assessment. Depending on the context and the verb used, compassion, approval, complaint or disapproval may be expressed:

(9) *–Kak ne ustaneš ty, njanja,* ***den'-den'skoj*** *deržat' na rukah Katjušu?* 'Nanny, how don't you get tired holding Katiusha all day long?' (RNC).

(10) *–I dvuh trudodnej ne zahočeš, kak pobudeš s nimi* ***den'-den'skoj****, a on skrjažničaet, volčij zub!* 'You won't desire even two workdays counted for one day's work, as you stay with them all day long, and he is skimping' (RNC, Sholokhov).

(11) *–Počemu nedelikatno? –* ***Den'-den'-skoj*** *sidit, ne vygoniš ego! Ja ego dnëm posylau guljat', a on na menja ogryazaetsja…* 'Why indelicate? He's sitting all day long; you can't kick him out! I send him to take a walk during the day, and he just snaps at me …' (RNC).

The lexicographical definition of the *den'-den'skoj* construction may look as follows*:*

Den'-Den'skoj P =

'(a) a situation **P takes place**;
(b) **P** lasts all day long;
(c) the speaker thinks that **P** takes a very long time'.

Such an interpretation explains why the microunit *den'-den'skoj* is not occurring outside of a durative construction.

At the same time, cognate *X X-ovyj* constructions have syntactic functions which are typical for noun phrases in sentences.

3 Syntactic Function

As already mentioned, the *den'-den'skoj* construction acts like an adverbial with a durative meaning. No phrases could be found in the Russian National Corpus or Araneum Russicum corpus with any elements dependent on the construction.

Apparently, *den'-den'skoj* is a predicate with a single valence denoting a durative process or state, so it is used in all contexts where such a situation can be specified by means of gerunds, habitual and repeated imperfective verbs, as well as many other ways, some of which will be discussed below. Perfective verbs rarely co-occur with the *den'-den'skoj* adverbial:

pojmal rybu* ***den'-den'skoj*, '*caught fish all day long' but *lovit rybu* ***den'-den'skoj*** 'fishing all day long'.

Nevertheless, the perfective aspect may avoid the restriction once the verb is of delimitative or perdurative Aktionsart (this works for all durative constructions):

(12) *Pobegaet* ***den'-den'skoj*** *po delam, a domoj vozvraŝaetsja zloj i ustavšij.* 'He runs errands all day long and comes back home angry and tired'.

(13) *On provaljalsja* ***den'-den'skoj*** *v posteli, tol'ko k noči prinjalsja za stat'ju.* 'He lay in bed all day long, and only at nightfall started to work on the article'.

The same applies to potential situations with a verb in future tense or dative subject constructions with an infinitive, like construction *Z-u X-ovat'* 'Z is to X':

(14) *Ne vsjakij vystoit* ***den'-den'skoj*** *za prilavkom.* 'Not everyone will stand behind the counter all day long'.
(15) *Ej li* ***den'-den'skoj*** *stojat' za prilavkom!* 'She should not have to stand behind the counter all day long!'.

We should admit that in the cases above, the semantic component of a usual and typical action is not implied due to the verbal Aktionsart. (12) and (13) refer to one-time situations.

The verbs combining with *den'-den'skoj* unit can roughly be distributed among the following three types:

1. Verbs denoting low mobility activities, monotonous and poorly controlled processes and states with a human subject, such as *sidet'* 'sit' and its derivates (the most frequently used in combinations), *spat'* 'sleep', *ležat'* 'lie', *valjat'sja* 'rest lying', *torčat'* 'hang around', *ždat'* 'wait', *smotret'* 'look', *revet'* 'cry', *pilit'* 'nag', *glazet'* 'stare', *rashaživat'* 'stroll', *katat'sja* 'ride', *perekladyvat'* 'shuffle' etc.
2. Verbs denoting intense labor or physical activity with an animate subject: *begat'* 'run', *hodit'* 'walk', *trudit'sja* 'labor', *rabotat'* 'work', *pahat'* 'work hard', *nosit'sja* 'scamper', *snovat'* 'scurry', *vertet'sja* 'spin', *igrat'* 'play', *rezvit'sja* 'frolic' etc.
3. Verbs denoting any durative action of natural phenomena: *razdavat'sja* (o zvone) 'ring (bell)', *šumet'* (o vetre) 'blow (wind)', *tjanut'sja* 'go on', *plyt'* (ob oblakah) 'float (clouds)', *stučat'* (o dožde) 'patter (rain)' etc.

The semantic components 'duration from the morning till the evening' and 'routine' are common to all cases, as is anthropomorphism: if no human observer is implied, the construction can hardly be used: **Na Venere* ***den'-den'skoj*** *žarko* 'It is hot on Venus all day long*'*.

With the verbs of the first group, *den'-den'skoj* adverbial can additionally communicate a disapproval or a complaint. It can serve as an indication to pointless, boring or idle activities, making the statement more expressive when describing a hateful job or a bummer:

(16) *V každom dome objazatel'no najdëtsja neskol'ko samozabvennyh spletnic, provodjaŝih ves'* ***den'-den'skoj*** *u pod"ezda.* 'In every house there are sure to be several selfless gossips who spend all day long sitting at the entrance.' (RNC).

With verbs of the second group, *den'-den'skoj* can introduce an empathy component. This happens in cases with the actant in the second or third person, but in cases of the first person, the empathy (with oneself) turns into a complaint:

Ona den'-den'skoj *na nogah, na minutku daže ne prisela.* 'She is spending all day long on her feet, she hasn't even sat down for a minute'.

vs

Ja den'-den'skoj *na nogah, na minutku daže ne prisela.* 'I spent all day long on my feet, I haven't even sat down for a minute.'

With verbs of the third group the attitude of the speaker is normally not expressed:

(17) *V zooparkah inogda ustraivajut ploŝadku dlja molodnjaka, na kotoroj samye raznye detënyši—ot kozlov i krolikov do lisjat I medvežat* ***den'-den'skoj*** *igrajut grug s drugom.* 'In a zoo, sometimes there is baby playground set up, where cubs, from little goats and rabbits to little foxes and bears, play all day long.' (RNC).

(18) *I neugomonno* ***den'-den'skoj*** *v golyh vetvjah berëz hlopotali belonosye grači.* 'And the white-necked rooks were tirelessly buzzing about in the bare branches of the birches all day long.'(RNC).

Unlike quasi-synonymous expressions like *ves' den'* 'all the day' or *dni naprolët* 'days long', 'from day to day', the syntactic idiom *den'-den'skoj* is difficult to negate, although any other component of the sentence can be negated:

On ej pesni poët* ***ne den'-den'skoj**.* '*He is singing songs to her not all day long'

On ej pesni ***ne poët*** *den'-den'skoj.* 'He is not singing songs to her all day long'

Ne ej *on pesni poët den'-den'skoj.* 'It's not her to whom he is singing his songs all days long'.

Considering the communicative structure of sentences in which this microsyntactic unit occurs, we should note that it cannot act as a topic. Even in cases of the initial position in a sentence, it is not a topic but a component of a rheme, like the adverb *davno* 'long ago' [15, 16]:

(19) ***Den'-den'skoj*** *ne umolkaet suhaja treskotnja kuznečikov.* 'The dry chirping of grasshoppers lasts all day long.'(RNC).

(20) ***Den'-den'skoj*** *topajuŝemu v lesu da v pole, na holode, na vetru stroevomu komandiru pitanie nužno bylo krepkoe.* 'The combat commander stomping in the forest and the field, in the cold wind all day long, needed good nutrition.' (RNC).

4 Conclusion

Russian repetitive expressions are widely used in colloquial speech and literature as a technique that makes an expression figurative, graphic, and emotional. There are at least three hundred constructions with repeated elements that may be regarded as microsyntactic units. The large number and great variety of such constructions indicate their importance and efficiency for conveying subtle meanings and emotions. Some of them are extremely frequent like *ele-ele* 'barely', *čut'- čut'* 'slightly', some of them are highly productive, like *X X-om, a / no* 'let X be X, but' (*dela delami, a* *sem'ja važnee* 'business is business but the family is more important'). The *Den'-den'skoj* unit considered in this paper is not very popular and at first glance may seem hardly remarkable. However, at a closer look, it reveals a unique set of very specific properties that distinguish it from other cognate constructions.

Acknowledgements The reported study was performed with the support of a grant from the Ministry of Science and Higher education of Russia, No. 075-15-2020-793 «New generation Computational Linguistics Platform for Digital Documentation of the Russian Language: infrastructure, resources, scientific research». The author is grateful to her teacher Leonid Iomdin for the support and important advice in preparing this paper.

References

1. Švedova, N. Ju. (1960). *Ocherki po sintaksisu russkoi razgovornoi rechi [Sketches on Syntax of Colloquial Russian Speech].* USSR Academy of Sciences Press. (in Russian).
2. Šmelëv, D. (1976). *Sintaksicheskaya chlenimost' vyskazyvaniya v sovremennom russkom yazyke.* Prosvescheniye. (in Russian).
3. Mel'čuk, I. (1987). Un affixe dérivationnel et un phrasème syntaxique du russe moderne: Essai de description formelle. *Revue des études slaves, 59*, 631–648.
4. Mel 'čuk, I. (1995). Phrasemes in language and phraseology in linguistics. In M. Everaert, E.-J. van der Linden, A. Schenk & R. Schreuder (Eds.), *Idioms: Structural and psychological perspectives* (pp. 167–232). Lawrence Erlbaum Associates.
5. Jackendoff, R. (1997). Twisting the night away. *Language, 73*, 534–559.
6. Apresjan, Ju., & Iomdin, L. (1989). The construction of the NEGDE SPAT' type: syntax, semantics, lexicography. [Konstrukcija tipa NEGDE SPAT': sintaksis, semantika, leksikografija]. Semiotika i informatika (pp. 34–92). Vsesojuznyj institut nauchnoj i texnicheskoj informacii, AN SSSR. (in Russian).
7. Apresjan, Ju., Boguslavsky, I., Iomdin, L., & Sannikov, V. (2010). Theoretical problems of Russian syntax. Interaction of the grammar and the lexicon. [Teoretičeskie problemy russkogo sintaksisa]. In: Apresjan, J. (ed). Jazyki slavjanskix kultur. (in Russian).
8. Iomdin, L. (2017). Between the syntactic idiom and syntactic construction. Nontrivial cases of microsyntactic ambiguity. [Meždu sintaksičeskoj frazemoj i sintaksičeskoj konstruktsiej. Netrivial'nye slučai mikrosintaksičeskoj neodnoznačnosti]. SLAVIA 2017, časopis pro slovanskou filologii, ročník 68, sešit 2–3 (pp. 230–243). Praha. (in Russian).
9. Iomdin, L. (2015). Konstruktsii mikrosintaksisa, obrazovannye russkoj leksemoj raz. [Constructions of microsyntax built by the Russian word raz.]. SLAVIA 2015, Časopis pro Slovanskou filologii, ročník 84, sešit 3 (pp. 291–30). Praha. (in Russian).
10. Avgustinova, T., & Iomdin, L. (2019). Towards a Typology of Microsyntactic Constructions. In G. Corpas Pastor & R. Mitkov (eds) Computational and corpus-based phraseology. EUROPHRAS (pp. 15–30). Springer.
11. Apresjan Ju. (2009). O proekte aktivnogo slovaria russkogo iazyka [About the project of Active dictionary of the Russian language] (vol. 3:56, pp. 118–130). Vestnik Rossiiskogo gumanitarnogo nauchnogo fonda [Vestnik of Russian Foundation for Humanities], Moscow. (in Russian).
12. The Russian National Corpus (ruscorpora.ru). 2003—2023. *Araneum Russicum corpus.* http://ucts.uniba.sk/aranea_about/_russicum.html
13. Baranov, A., & Dobrovol'skij, D. (2008). Aspects of phraseological theory. Znak. (in Russian).
14. Apresjan, Ju. (1992). Leksikografičeskie portrety (na primere glagola byt') [Lexicographic portraits (illustrated by the verb BYT' 'be')]. *Naučno-texničeskaja informacija, 3*, 20–33.

15. Yanko, T. (2001). *Communicative strategies of Russian speech [Kommunikativniye strategii russkoy rechi]*. Yazyki slavyanskoy kul'tury. (in Russian).
16. Padučeva, E. (1996). Davno i dolgo. Logicheskij analiz Semantika vremeni i vida v russkom jazyke. Jazyk I vremya. Indrik (pp. 253–266). (in Russian).
17. Nunberg, G., & Sag, I. A. (1994). Thomas Wasow.: Idioms. *Language 70*(3), 491–538.

“Plain and Natural” Versus “Accurate and Unambiguous”: Pronominal Intrasentential Anaphora in Russian Legislative Texts

Olga Blinova

1 Introduction

The paper focuses on Russian legislative texts, more precisely, on some peculiarities of referential cohesion in such texts. The paper’s aim is a corpus-based description of reduced referential devices’ usage.

An anaphoric expression is a referential device which an author uses to refer to the previously mentioned entity. It is customary to distinguish between complete referential expressions (semantically complete noun phrases) and reduced referential expressions (anaphoric pronouns and zeros). In the Russian language, personal, reflexive, demonstrative and relative pronouns are capable of performing anaphoric functions.

I will focus on describing 3rd person personal pronouns in anaphoric function. Researchers of legal English tend to agree that “legal drafters have traditionally avoided using personal pronouns <…> . The reason for this is a fear of ambiguity in cases where it is unclear to which noun the pronoun might refer” [1, p. 5], see also [2, p. 88], [3, p. 113].

The authors of legal texts in various languages avoid personal pronouns. Because of this, there is a large number of repetitions of complete noun phrases, which turns legal texts into “formal and intimidating” [1, p. 5]. Repetitions make the texts unnatural, formalized, and, according to one possible view on the problem, more complex, see [4].

Thus, legal drafters are forced to find a balance between naturalness and simplicity on the one hand, and legal precision and referential unambiguity on the other.

O. Blinova (✉)
St. Petersburg University, Universitetskaya Emb. 7/9, 199034 St. Petersburg, Russia
e-mail: o.blinova@spbu.ru; ovblinova@hse.ru

National Research University Higher School of Economics, 16, Ul. Soyuza Pechatnikov, 190068 St. Petersburg, Russia

P. Eismont et al. (eds.), *Literature, Language and Computing*,
https://doi.org/10.1007/978-981-99-3604-5_3

Table 1 The composition of the CorCodex corpus

Document type	N of documents
Federal Law	199
Government Resolution	37
Russian Federation Code	21
Law of the Russian Federation	17
Fundamentals of Russian legislation	2
Resolution of the Supreme Council	2
Federal Constitutional Law	1

2 Material and Method

The source of language data is the CorCodex legal synchronous corpus. It consists of 279 documents issued from 1991 to 2020. This corpus includes codes, federal laws of the Russian Federation, and other regulatory documents. The corpus size is 3 million 227 thousand tokens, 2 million 722 thousand words.

The CorCodex corpus has been lemmatized, morphologically and syntactically annotated using UDPipe (the "ru-syntagrus" model) and pymorphy2 [5, 6], and published on the website https://www.plaindocument.org/. The corpus composition is described in Table 1.

Part-of-speech tagging in terms of Universal Dependencies makes it possible to distinguish between PRON (mostly noun pronouns; personal pronouns have been marked with this tag) and DET (mostly adjective pronouns).

The sample included in the analysis consists of sentences with singular uses of 3rd person personal pronouns, cf. example (1).

(1) ***Kreditnaja organizacija*** *ne vprave osushhestvljat' jemissiju obligacij s ipotechnym pokrytiem, esli* ***ona*** *ne vypolnjaet hotja by odno iz trebova-nij, ustanovlennyh v sootvetstvii s polozhenijami nastojashhej stat'i.* [CorCodex, Federal Law #152, 2003].

A credit institution is not entitled to issue mortgage-backed bonds if **it** does not fulfill at least one of the requirements established in accordance with the provisions of this article

The corpus consists of 139,285 sentences. 23,816 sentences contain PRON or DET usage one or more times. This study analyzes a sample of 1,917 sentences containing a single use of 3P personal pronoun.

The model of referential choice [7, 8] was used as a description scheme, since it has been repeatedly tested on Russian language data in corpus-based studies.

The sample of sentences has been annotated according to the following parameters[1]:

1. the type of referent,
2. the length of antecedent in words,
3. the type of syntactic phrase of antecedent,
4. linear distance between antecedent and anaphor.

In addition, the analysis involved calculating examples in which referential expressions (the anaphora and its antecedent) are not within the boundaries of a single sentence, that is, calculating cases of discourse anaphora, cf. (2).

(2) *Porjadok prohozhdenija* ***imi*** **<**they-INS**>** *voennoj sluzhby v Sledstvennom komitete reguliruetsja Federal’nym za-konom ot 28 marta 1998 goda N 53-FZ* <...> . [CorCodex, Federal Law #403, 2010].

The procedure for **their** military service in the Investigative Committee is regulated by Federal Law No. 53-FZ of March 28, 1998 <…>

This paper presents some results of a quantitative analysis of a sample annotated for all of the above-mentioned parameters.

3 Results

3.1 Discourse and Sentential Anaphora

The case in which the meaning of anaphor cannot be interpreted within the boundaries of a sentence is not frequent in the sample. There are a total of 26 instances of discourse anaphora per 1,917 sentences analyzed, which is about 1.36%. Thus, 98.64% of the anaphors are placed in the same sentence as the antecedents. According to [9], in English 90% of antecedents are placed in the same sentence as their pronominal anaphors. If one uses [Ibid., p. 323] to formulate theoretical expectations, about 10% of the examples could come from cases of discourse anaphora. Thus, one can conclude that the drafters of Russian legal texts strive to ensure that the antecedent and anaphor are within the same sentence.

When considering the distribution of cases of discourse anaphora by year (see Fig. 1), one can notice an increase in the frequency of corresponding instances in the early 2000s, but these observations are insignificant, as the number of uses is negligible.

[1] The preliminary results of the study performed on a smaller sample were previously presented by O. Blinova and Yu. Alekseeva in the paper “Personal pronoun as a reduced referential device in Russian legal text” at the conference “Russian language issues in legal cases and procedures” (May 18, 2021), as well as in the bachelor thesis by Yu. Alekseeva “Reduced referential devices in modern Russian legal documents (according to corpus data)” (2021).

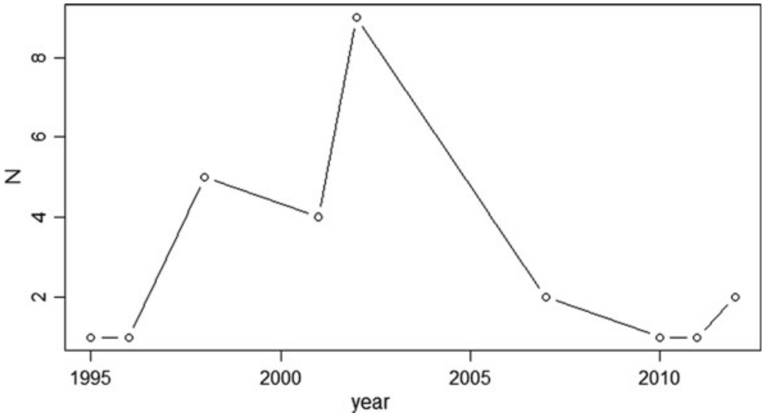


Fig. 1 Instances of discourse anaphora (by year)

In Sects. 3.2–3.5, only the cases of intrasentential anaphora are considered; the number of examples with discourse anaphora is insufficient for quantitative analysis.

3.2 Type of Referent

The three types of referents have been distinguished, namely animate, inanimate and collective ones, the latest refer to an organization or organizations, or both persons and organizations, see (3).

(3) L'goty po uplate tamozhennyh platezhej predostavljajutsja ***inostrannym investoram i kommercheskim organizacijam s inostrannymi investicijami*** *pri osushhestvlenii* ***imi*** <they-INS> *prioritetnogo investicionnogo proekta v sootvetstvii s tamozhennym zakonodatel'stvom Rossijskoj Federacii i zakonodatel'stvom Rossijskoj Federacii o nalogah i sborah.* [CorCodex, Federal Law #160, 1999].

> Benefits for payment of customs duties are granted to **foreign investors and commercial organizations with foreign investment** in the implementation of **their** priority investment project in accordance with the customs legislation of the Russian Federation and the legislation of the Russian Federation on taxes and fees.

It turned out that in sentences with intrasentential anaphora inanimate referents predominate (see Table 2).

Table 2 Types of referents

Type of referent	Number	Percent
Inanimate	977	51.67
Animate	667	35.27
Collective	247	13.06

Pronominal anaphoric expressions refer to the “most accessible” entities that are already known, or “given” to an addressee. Such entities are denoted in theories of referentiality as “familiar”, “in focus”, “accessible”, “(highly) activated” ones, see, inter alia, [10]. Referent’s animacy/inanimacy is able to affect its activation. For example, [11, p. 55] shows that “animacy and the status of the protagonist are able to increase the activation coefficient in the case of reactivation of the referent, that is, at relatively large distances to the antecedent”. Thus, the data on observed referent types is not enough; one has to use information about how the referent type and the linear distance between the anaphor and the antecedent correlate, which may be a research perspective.

3.3 Length of Antecedent

The length of the antecedent is also traditionally considered in referential models. In the described sample, the length of the antecedent varies from 1 to 131 words, but the value of 131 is an outlier (see Fig. 2). The median antecedent length is 2 words, and the mean antecedent length is 3.23 words.

This parameter can be useful for showing the differences between legal and, more generally, official texts, and texts of other styles. The differences can be explicated by operating on the values of the average length of sentences, phrases, words. In addition, the parameter is related to the one considered in paragraph 3.4, since the more noun phrases there are in a coordinated phrase, the longer it is.

Finally, the length of the referential expression correlates with the degree of activation of the corresponding referent (the longer the expression, the less activated the referent), cf. the hierarchy of accessibility markers in [12, p. 449].

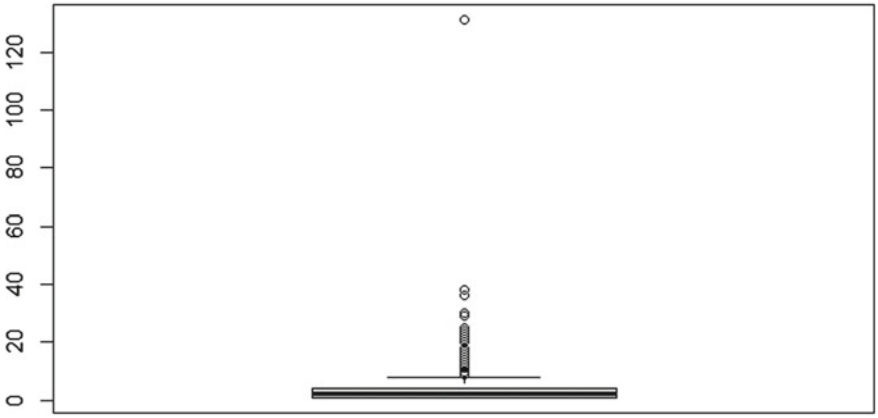


Fig. 2 The length of antecedent in words

3.4 Type of Antecedent

The distinction between NP (noun phrase) and CNP (coordinated noun phrase) has been used to annotate the types of syntactic phrases of antecedents. Information on the number of NPs in CNPs has not been taken into account. The length of CNP in the sample is up to 10 NPs, cf. (4).

(4) Zakonodatel'stvo Rossijskoj Federacii o nakopitel'noj pensii sostoit iz **nastojashhego Federal'nogo zakona, Federal'nogo zakona ot 16 ijulja 1999 goda N 165-FZ "Ob osnovah objazatel'nogo social'nogo strahovanija", Federal'nogo zakona ot 15 dekabrja 2001 goda N 167-FZ "Ob objazatel'nom pensionnom strahovanii v Rossijskoj Federacii", Federal'nogo zakona ot 24 ijulja 2009 goda N 212-FZ "O strahovyh vznosah v Pensionnyj fond Rossijskoj Federacii, Fond social'nogo strahovanija Rossijskoj Federacii, Federal'nyj fond objazatel'nogo medicinskogo strahovanija", Federal'nogo zakona ot 7 maja 1998 goda N 75-FZ "O negosudarstvennyh pensionnyh fondah", Federal'nogo zakona ot 30 nojabrja 2011 goda N 360-FZ "O porjadke finansirovanija vyplat za schet sredstv pensionnyh nakoplenij", Federal'nogo zakona ot 1 aprelja 1996 goda N 27-FZ "Ob individual'nom (personificirovannom) uchete v sisteme objazatel'nogo pensionnogo strahovanija", Federal'nogo zakona ot 24 ijulja 2002 goda N 111-FZ "Ob investirovanii sredstv dlja finansirovanija nakopitel'noj pensii v Rossijskoj Federacii", drugih federal'nyh zakonov** i prinimaemyh v sootvetstvii s **nimi** normativnyh pravovyh aktov Rossijskoj Federacii. [CorCodex, Federal Law #424, 2013].

> The legislation of the Russian Federation on funded pensions consists of **this Federal Law, Federal Law of July 16, 1999 N 165-FZ "On the Basics of Compulsory Social Insurance", Federal Law of December 15, 2001 N 167-FZ "On Compulsory Pension Insurance in the Russian Federation ", Federal Law of July 24, 2009 N 212-FZ "On insurance contributions to the Pension Fund of the Russian Federation, the Social Insurance Fund of the Russian Federation, the Federal Compulsory Medical Insurance Fund", Federal Law of May 7, 1998 N 75-FZ "On non-state pension funds", Federal Law of November 30, 2011 N 360-FZ "On the procedure for financing payments from pension savings", Federal Law of April 1, 1996 N 27-FZ "On individual (personalized) accounting in the system of compulsory pension insurance", Federal Law of July 24, 2002 N 111-FZ "On investing funds for financing funded pension in the Russian Federation", other federal laws** and regulatory legal acts of the Russian Federation adopted in accordance with **them**.

In Russian legal texts, there are specific "double" coordinating conjunctions *i (ili)* 'and (or)' suggesting an alternative conjunctive vs disjunctive interpretation, see (5). In the sample of CNP-type examples there are 8 entries with such double conjunctions, which is 0.4%.

(5) *Strategicheskie predprijatie i organizacija schitajutsja nesposobnymi udovletvorit' trebovanija kreditorov po denezhnym objazatel'stvam i (ili) ispolnit' objazannost' po uplate objazatel'nyh platezhej, esli* ***sootvetstvujushhie objazatel'stva i (ili) objazannosti*** *ne ispolneny v techenie shesti mesjacev s daty, kogda* ***oni*** *dolzhny byli byt' ispolneny.* [CorCodex, Federal Law #127, 2002].

Table 3 Types of syntactic phrases of antecedents

Type	Number	Percent
NP	1639	86.67
CNP	252	13.33

> 'A strategic enterprise and organization is considered unable to satisfy the claims of creditors on monetary obligations and (or) fulfill the obligation to pay obligatory payments, if **the relevant liabilities and (or) obligations** are not fulfilled within six months from the date when **they** should have been fulfilled.'

The information about the number of NP-type and CNP-type antecedents in the annotated sample is presented in Table 3.

The parameter described is also intended to highlight the specifics of legal texts containing series of coordinated constituents.

3.5 Distance between Antecedent and Anaphor

Traditionally, the models of reference take into account linear distance between antecedent and anaphor either as a basic factor or as an auxiliary one. It is stated that the greater the distance, the less the degree of accessibility of the referent (and, accordingly, the less likely it is that a reduced referential device will be used when the referent is mentioned again), see, for example, [13, p. 446–447]. The distance can be calculated in clauses or in words. In addition, in a written text one can consider paragraph boundaries.

The following solution has been used to annotate the sample. When an antecedent and an anaphor were within the same clause, such examples have received the index "0"; if they were in neighboring clauses, such occurrences have been assigned the index "1", if an antecedent and an anaphor were separated by one clause, the index "2" has been used, etc. The distribution of distances in clauses is shown in Table 4.

Table 4 Linear distance (in clauses)

Index	Number	Percent
0	1166	61.66
1	581	30.72
2	97	5.13
3	35	1.85
4	3	0.16
5	3	0.16
6	3	0.16
7	1	0.05
9	2	0.11

The largest observed value of distance is 9 clauses. However, the cumulative proportion of examples with the distance exceeding two clauses is 2.5%.

Table 5 shows the values of distance between the anaphor and antecedent in words.

The mean value for the word distance is 7.04, the median value is equal to 5. Looking at Fig. 3, it is clear that the values of distances above18 are outliers, and values like “86”, “91”, “115”, and “141” are very far from the median value.

When interpreting linear distance values, it is worth bearing in mind that, according to [14], “regular pronouns normally tend not to refer to antecedents inside the same clause”; that is, there is a tendency to place the anaphor and the antecedent in different clauses (except for reflexive and reciprocal pronouns). In the observed data, the anaphor and antecedent are located in the same clause in 61.66% of occurrences. This may be due to the fact that the average length of clauses in legal texts is higher than in other texts in Russian.

Table 5 Linear distance (in words)

Distance	Number	Distance	Number
1	162	25	7
2	254	26	4
3	267	27	5
4	206	28	3
5	210	29	3
6	115	30	5
7	90	31	1
8	72	32	6
9	92	33	2
10	60	34	1
11	52	35	1
12	46	36	1
13	39	37	2
14	32	39	2
15	28	42	1
16	22	43	1
17	22	49	1
18	17	55	1
19	10	60	1
20	7	86	1
21	9	91	1
22	6	115	1
23	14	141	1
24	7		

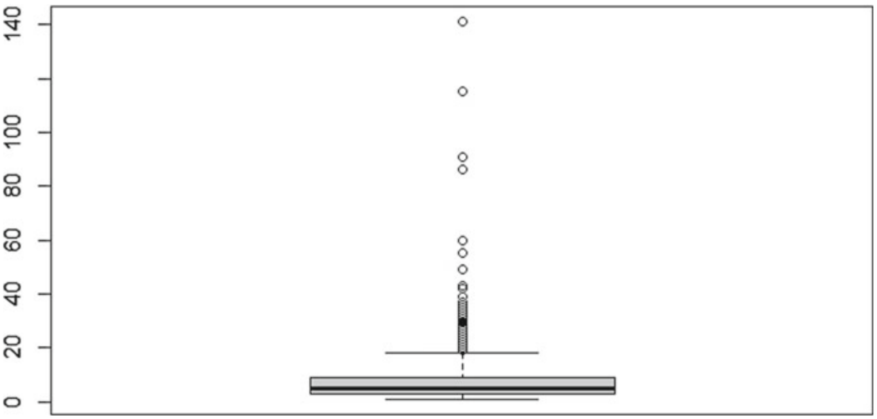


Fig. 3 Linear distance (in words)

Let’s look at the sentence length values, related to the length of clauses. For the subset of sentences with intra-sentential pronominal anaphora, the maximum sentence length in words is equal to 292, the average sentence length (ASL) is 34.48 words, and the median sentence length is 29 words, see Fig. 4. For comparison, one can take school textbook data [15], where ASL value does not exceed 16.19 words. In [16] the cited ASL value is equal to 17.12 words. “For the Russian language in general” (if one considers SinTagRus corpus data representative) the ASL value is equal to 14.26 [17]. Thus, the ASL value for a legislative text is about twice as big as the general-language ASL value.

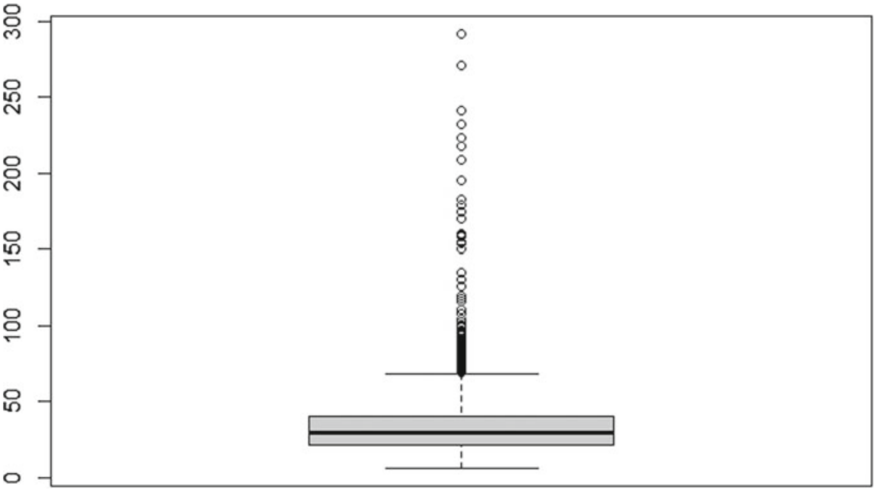


Fig. 4 Sentence length

4 Conclusion

This paper has analyzed the instances of pronominal anaphora in a legislative text. The sentences with single uses of 3P personal pronouns (1,917 sentences in total) have been considered. A set of description parameters was as follows: the type of referent, the length of antecedent in words, the type of syntactic phrase of antecedent, the linear distance between antecedent and anaphor.

It is noteworthy that the proportion of discourse anaphora instances is about 1.36%, which means that the authors tend to have the antecedent and the anaphor within the same sentence. In addition, according to the data obtained, 61.66% of the pronominal anaphors are located with their antecedent in the same clause, see examples (6) and (7).

(6) *Po pros'be* ***sotrudnika*** *vmesto predostavlenija dopolnitel'nyh dnej otdyha* ***emu*** *mozhet byt' vyplachena denezhnaja kompensacija.* [CorCodex, Federal Law #197, 2018].

At the request of **an employee**, **he** may be paid monetary compensation instead of additional days of rest.

(7) ***Vopros o prinjatii zajavlenija ili predstavlenija k proizvodstvu*** *rassmatrivaetsja edinolichno sud'ej Vysshego Arbitrazhnogo Suda Rossijskoj Federacii v pjatidnevnyj srok so dnja* ***ego*** *postuplenija v Vysshij Arbitrazhnyj Sud Rossijskoj Federacii.* [CorCodex, Arbitration Procedural Code, 2002].

The question of admitting an application or representation to proceedings shall be considered by a single judge of the Supreme Arbitration Court of the Russian Federation within five days from the date of **its** receipt by the Supreme Arbitration Court of the Russian Federation.

This can be explained by the legal drafters' fear of generating referentially ambiguous and difficult-to-interpret passages and their intention to "keep the antecedent and the anaphor closer to each other". Indeed, the greater the distance from the antecedent, the more difficult it is to interpret the anaphoric pronoun (and the lower the coherence of the text). The average value of referential distance in words in the sample is 7.04, with a median value of 5.

"Keeping the antecedent and anaphor close to each other" turns out to be difficult, since syntactic units in the sample are, on average, significantly longer than "in the Russian language in general", e. g. the ASL (average sentence length) value is equal to 34.48 words, while in SinTagRus corpus [17] the value of this measure is 14.26.[1–3, 11–13]

Acknowledgements The research was supported by the Russian Science Foundation, project #19-18-00525 "Understanding official Russian: the legal and linguistic issues".

References

1. Haigh, R. (2004). *Writing legal English.* London, Sydney, Portland: Cavendish Publishing Ltd.
2. Mattila, H. E. S. (2013). Comparative legal linguistics: Language of law, Latin and modern Lingua Francas.
3. Williams, C. (2004). Legal English and plain language: An introduction. *ESP Across Cultures, 1*, 111–124.
4. Saveliev, D. A., & Kuchakov, R. K. (2019). *Decisions of arbitration courts of Russian Federation: Lexical and syntactic quality of texts, analytic note.* St. Petersburg: European University at Saint Peters-burg. [in Russian].
5. Universal Dependencies 2.5. Models for UDPipe. https://github.com/jwijffels/udpipe.models.ud.2.5/blob/master/inst/udpipe-ud-2.5-191206. Accessed 21 May 2022.
6. Korobov, M. (2015). Morphological analyzer and generator for Russian and Ukrainian languages. *Analysis of images, social networks and texts* (pp. 320–332).
7. Kibrik, A. A. (1999). Cognitive inferences from discourse observations. In K. Van Hoek, A.A. Kibrik & L. Noordman (Eds.), *Discourse studies in cognitive linguistics: Selected papers from the 5th International cognitive linguistics conference* [Current Issues in Linguistic Theory 176] (pp. 29–52).
8. Kibrik, A. A., Dobrov, G. B., Zalmanov, D. A., Linnik, A. S., & Lukashevich, N. V. (2010). Referential choice as a multifactorial probabilistic process. In: Komp’juternaja Lingvistika i Intellektual’nye Tehnologii (vol. 9, no. 16, pp. 173–180) [in Russian].
9. Hobbs, J. R. (1987). Resolving pronoun references. *Lingua, 44*, 311–338.
10. Navarretta, C. (2005). Combining centering-based models of salience and information structure for resolving intersentential pronominal anaphora. In A. Branco, A. M. McEnery, & R. Mitkov (Eds.), *Anaphora processing: Linguistic, cognitive, and computational modelling: Selected papers from DAARC 2002* (pp. 329–350). Amsterdam: John Benjamins Publishing Co.
11. Kibrik, A. A. (2003). *Discourse analysis in a cognitive perspective, dissertation of the doctor philology*. Moscow: Institute of Linguistics RAS.
12. Ariel, M. (2001). Accessibility theory: An overview. In T. Sanders, J. Schilperoord, & W. Spooren (Eds.), *Text representation: Linguistic and psycholinguistic aspects* (pp. 29–87). Amsterdam: John Benjamins Publishing Company.
13. Ariel, M. (1991). The function of accessibility in a theory of grammar. *Journal of Pragmatics, 16*(5), 443–463.
14. Kibrik, A. A. (2004). Review of: Anaphora: A cross-linguistic study, by Yan Huang. Oxford: OUP, 2000. *Linguistic Typology, 8*(3), 389–393.
15. Solnyshkina, M., Solovyev, V., Ivanov, V., & Danilov, A. (2018). Studying text complexity in Russian academic corpus with multi-level annotation. In *Computational models in language and speech workshop* (CMLS 2018) (vol. 2303, pp. 93–103).
16. Ivanov, V. V., Solnyshkina, M. I., & Solovyev, V. D. (2018). Efficiency of text readability features in Russian academic texts. In *Komp’juternaja Lingvistika i Intellektual’nye Tehnologii* (vol. 17, pp. 277–287).
17. Zeman, D., Dušek, O., Mareček, D., Popel, M., Ramasamy, L., Štěpánek, J., Žabokrtský, Z., & Hajič, J. (2014). HamleDT: Harmonized multi-language dependency treebank. *Language Resources and Evaluation, 48*(4), 601–637.

The Old Church Slavonic Corpora and Their Use in Language Studies at the University

Ilia Afanasev and Andrey Babanov

Abstract The paper focuses on the issues that surround the research of Old Church Slavonic (OCS). The most important point is that, despite the number of existing OCS corpora, none of them actually fully represents the language. Some of them contain texts that are not OCS ones, they are merely Old and Slavic. And each of these corpora is highly centralized, which makes them extremely vulnerable. The paper discusses corpora preservation issues that are becoming increasingly important for OCS studies. Another problem is that the OCS canon is extremely heterogeneous, and it is hard to discriminate the OCS texts from other Slavic texts of the approximately same time. The possible formal criteria are discussed in the article, as well as a set of texts that correspond to these criteria and should form the OCS corpus. The process of corpus construction is also addressed in the paper. The emphasis is on the three modules. The first helps to tag the data or edit it after automatic tagging. The second helps to visualize the relationships within the data, and the third is a corpus search module. Finally, the ways in which the corpus created can be used in the educational process are given.

Keywords Old Church Slavonic · Historical linguistics · Corpus linguistics · Applied linguistics · Dialect variation

I. Afanasev (✉)
National Research University Higher School of Economics, 20 Myasnitskaya Ulitsa, 101000 Moscow, Russia
e-mail: ilia.afanasev.1997@gmail.com

MTS Artificial Intelligence Center, LLC, 18 Andropova Ave., build. 9, off. 21, fl.3, 115432 Moscow, Russia

A. Babanov
St. Petersburg University, Universitetskaya Emb. 7/9, 199034 St. Petersburg, Russia
e-mail: a.babanov@spbu.ru

P. Eismont et al. (eds.), *Literature, Language and Computing*,
https://doi.org/10.1007/978-981-99-3604-5_4

1 Introduction

The Old Church Slavonic (OCS) corpora are vulnerable. There are a lot of them now [6, 16, 17] , however, each of them may fail at any given moment of time. Creating an extra may seem futile, but providing the corpus that is easily distributed and decentralized is. Also, it should be taken into consideration that existing OCS resources need extension and reinforcement.

It is even of greater importance as recent years have witnessed the emergence of the OCS research that has utilized manually collected texts. The exemplary work is the corpus-based study of the OCS verb aspect [13]. The OCS studies currently witness the growing necessity for corpus-based research, combined with a lack of corpora that are able to satisfy this necessity. For Slavic Studies this is a developing branch, while in corpus linguistics with each year it needs more and more revitalization.

The corpora creation also provides an important exercise of reevaluating all the OCS canon texts as representing exactly this language. This is because the OCS canon is characterized by an enormous degree of linguistic heterogeneity [27]. Texts should be picked according to narrowly defined, formalized linguistic criteria. The current article uses the criteria given in different theoretical works, aimed at the OCS study, for instance, in [13]. And not all the texts, previously defined as OCS, are going to comply with these criteria.

Once the corpus is prepared, it may be used for different purposes [4]. The tutorial programs may be prepared for the training of word under *titlo* comprehension, a graphical variety of word forms and its reasons for understanding, systematization and understanding of semantic and syntactical word functions, providing students with the tools to develop their skills of working with Glagolitic texts. Thus, the corpus may become a modern tool for the versatile study of OCS both as a *lang* and a *parole* [6].

The work is structured as follows. Section 2 provides the background on the OCS corpora which currently exist. Section 3 gives the analysis of the OCS texts by the formally defined criteria of their belonging to this canon, as well as the theoretical background. Section 4 briefly suggests how the corpus can be used in university language studies. Section 5 provides a conclusion to the article.

2 The Actual OCS Corpora and Issues of Corpora Creation for OCS

The most representative text collection is presented by the *Corpus Cyrillo-Methodiana Helsinginiense* of the University of Helsinki and contains *Codex Assemanius, Codex Marianus, Codex Suprasliensis, Codex Zographiensis, Liber Sabbae*, and (from the later copies) *Vita Constantinii* and *Vita Methodii* [6]. The corpus is part of the Frankfurt Goethe University text collection that also incorporates *Kyiv Folia* (traditionally called *Kiev Leaflets*) and *Prague Fragments*, the manuscript that traditionally is considered to be the oldest Czech Church Slavonic, being placed between the two stages of the single language system development. It seems that it was included as an OCS text because it contained more features of the ancestor idiom, than those of the descendant idiom, and was the thing that is usually called transitional fossil [9]. The exact reason is not mentioned by the researchers who made the collection, and the linguistic comparison of Prague Fragments and the OCS texts is to be the subject of the forthcoming research.

Both the *Corpus Cyrillo-Methodiana Helsinginiense* and the Goethe University text collection do not contain a significant number of manuscripts of different size, with each of these manuscripts, such as *Psalterium Sinaiticum*, being actively used in present-day research as part of the OCS canon [13]. It appears necessary to enrich the corpus by these manuscripts.

Similar fallacies also characterize the web application that is developed at the University of Oslo and represents the New Testament multi-language corpus, that includes the OCS texts as well [17]. The latter part has considerably grown in recent years, including *Codex Marianus, Codex Zographiensis, Codex Suprasliensis, Psalterium Sinaiticum, Vita Constantinii, Vita Methodii, Euchologium Sinaiticum*, and *Kyiv Folia.*

These texts are partially morphologically tagged. It is performed manually, it is above the average tagging level in terms of precision, though it is not scalable, and tagging of other texts would have to be done from scratch, require too many researchers and too much time. However, the corpus itself does not comply with the criteria of being full, representative and suitable for the full-scale research of OCS.

Manuscript project currently contains approximately a dozen of manuscripts that belong to the OCS canon, as stated by some researchers [19]. They are tagged and lemmatized, they have already become the material for some statistical linguistic research [19]. However, they are available on demand and are often fragmentary [19]. Moreover, a huge chunk of the OCS canon is yet to be incorporated into *Manuscript*. The project itself is aimed at the study of the ancient Slavonic literacy sensu lato, and does not aim at the creation of the tool for the interaction with namely OCS texts. Thus, it is not a fully complete and representative OCS corpus.

The other corpora of OCS, created over the past decades, are now either closed access, or disappeared from the Internet. The question of whether to use them or not does not make any sense by now.

There is also a historic corpus of the Bulgarian language [7]. Its texts typologically are very similar to the OCS canon. However, not a single one of these texts is traditionally defined as an OCS one, so at present it is not possible to call the corpus itself OCS.

3 The OCS Corpus Creation: The Experience of Text Choice

The main issue of the creation of an OCS corpus, is the spectrum of texts to be included. Due to the linguistic heterogeneity of OCS [29], any decision whether to include or not to include a text will be subjective, based on the set of criteria, some of which are not linguistic sensu stricto, more historical.

The latter is an especially important factor, when one takes into consideration the text of the times of Second Heimatland [13], *Kyiv Folia*, also known as *Kyiv Fragments*, or *Kyiv Missale*. "Of the times" is the key word here: *Kyiv Folia* were not written as part of Cyril and Methodius activity in the Great Moravia, but were created simultaneously. As scholars state [13], *Kyiv Folia* demonstrate both South Slavic and West Slavic idiom features, when the most important reflexes are considered, the reflexes that primarily are used for attributing the text to OCS, or one of the latter Church Slavonic languages. However, *Kyiv Folia* are included into the OCS canon by the majority of the researchers, who are analyzing its features [13, 31]. It makes the inclusion of *Kyiv Folia* into the OCS corpus necessary by the criterion of the scholar historical agreement. This criterion, thus, is not just the key one, but the one that is more important than the sum of all the other criteria.

In the future, when the corpus is digitalized, a more precise analysis for homo/heterogeneity will be possible, therefore, it may be divided into subcorpora. Or, probably, one may also redefine the whole definition of OCS using the new data. But as this time has not come yet, the defining criterion is the scholar historical agreement.

However, this criterion, even if being obviously extralinguistic, should not be totally arbitrary, when every OCS scholar may just claim that a text is definitely an OCS one. Not a single work, even perfectly methodologically competent, even a classic one, may in itself define, which text should be included in the corpus, and which should not. Thus, the initial decision should be made with a large number of works in mind. Each of these works should be a classical fundamental study, acknowledged by the community, or the existing OCS text collection or corpus, or the modern general work with a strong summarizing component. Later, after the discussion and corpus analysis, the set of texts may probably be changed, if necessary.

When creating the OCS corpus, we use the following body of works:

1. The classic OCS research, the most detailed works that have been written during the history of study of OCS. Mostly these are the works that have been written in the mid-twentieth century [20, 25, 30].
2. The category of modern fundamental works consists of, first of all, [13], which sets up a thorough investigation into the characteristics of the OCS texts and their key features by the group of criteria, both historical and linguistic, including **tj/*dj* [13].
3. The OCS dictionary that represents summarization of this language study in the period up to the end of the twentieth century and includes a list and thorough

description of manuscripts should be mentioned as a separate and important source [32].

4. The last category of works that should be pointed out is the OCS existing corpora, both annotated and not, such as *Corpus Cyrillo-Methodianum Helsinginiense* [6], built on its base part of the New Testament corpus [17], and *Thesaurus Indogermanischer Texten- und Sprachmaterialen* [16].

When combined, these works do not yet provide the full image, however, their, so to write, agreement, due to the analysis quality of each of them, may pretend to be sufficient. The possibility to include other works may be the subject to further discussion.

All the works that are in this sampling, are treated as equally important for making a decision whether to include or not to include a text in the OCS corpus. The alternative decision would make possible the further arbitrariness of the criterion, what, given the arguments above, should not be allowed.

According to this criterion, the texts are going to be split into three categories.

The first would include the texts that are discussed in all the works. Agreement on their issue has existed for more than seventy years, the scholars who create the corpora consider it obligatory to innclude them into the corpus. Putting these texts into the created corpus, this way, is making it minimally representative.

The second category would be presented by the texts that are in the works of at least three of the types mentioned above. They also naturally would be added into the corpus because wide agreement on their matter has been present for a long time. However, this addition would follow the addition of the first text group.

The third category would include the texts that are present in one or two types of the works mentioned above. They may be included in the OCS corpus on the basis of linguistic analysis, search for the unique OCS features, mainly, the **tj/*dj* reflex.

The texts are distributed as follows.

1.1. The texts that are found in 4 types of sources: *Codex Marianus* [6, 13, 16, 17, 20, 25, 30, 32], *Kyiv Folia* [6, 13, 16, 17, 20, 25, 30, 32], *Codex Zographiensis* [6, 13, 16, 20, 25, 30, 32], *Codex Assemanius* [6, 13, 16, 20, 25, 30, 32], *Liber Sabbae* [6, 13, 16, 20, 25, 30, 32], *Codex Suprasliensis* [6, 13, 16, 17, 20, 25, 30, 32].

1.2. The texts that are found in 3 types of sources: *Glagolita Closiana* [13, 20, 25, 30, 32], *Psalterium Sinaiticum* [13, 20, 25, 30, 32] *Euchologium Sinaiticum* [13, 20, 25, 30, 32], *Evangelium Achridanum* [13, 20, 25, 30, 32], *Hilandar Fragments* [13, 20, 25, 30, 32], *Undolski Fragments* [13, 20, 25, 30, 32], *Zograph Fragments* [13, 20, 25, 30, 32], *Rila Glagolitic Fragments* [13, 20, 25, 30, 32].

1.3. The texts that are found in less than 3 types of sources: *Samuil Inscription* [13, 20, 30], *Preslav Inscription* [13, 30], *Macedonian Fragment* [13, 30], *Martyrologium Odonis* [30], *Ostromir Gospels* [20, 25], *Nikolje Gospels* [20], *Vatican Palimpsest* [13], *Bojana Palimpsest* [13, 32], *Sinai Fragment* [13], *Enina Apostolos* [13, 32], *Psalterium Demetrii Sinaitici* [13], *Missale Sinaiticum* [13], *Saint Petersburg Octoechos* [13], *Zograph Palimpsest* [32], *Codex Zographiensis—B* [6, 16], *Vita Constantinii* [6, 16], *Vita Methodii* [6, 16], *Prague Fragments* [16].

The work [30] also mentions some texts that contain features of other Slavic languages, however, the author considers them to be comparative material rather than material to study. The only exception is *Kyiv Folia* that demonstrate, according to the author, archaism of enough degree to be considered the OCS ones [30].

The work [20] insists on addition of the whole group of texts that have features of other Slavic languages into the OCS ones, namely *Ostromir* and *Nikolje Gospels*. The author explains their decision in a predominantly non-linguistic way, making stress on "archaism and special importance" [20]. *Kyiv Folia* are straightforwardly called a Moravian Church Slavonic manuscript, and despite that they are named the OCS one [19].

The work [25] includes *Codex Suprasliensis* and *Kyiv Folia* into the OCS canon,, with some nuances, though. Firstly, it is stated that the influence of Moravian dialects on *Kyiv Folia* is significant, yet the secondary nature of this manuscript allows not to take its linguistic features into consideration while creating the whole image of OCS lexis and grammar. Secondly, *Codex Suprasliensis* does not have a significant number of archaic features, traditionally attributed to the "language of the first translators", however, the features themselves are not stated [25]. Despite all that, these features of these manuscripts are not considered to be the reason for not including them into the OCS canon.

The work [13] also mentions *Novgorod Fragments* and *Psalterium Slucae*, used as comparative material, but not considered as part of the OCS canon. In addition, the author points out that *Kyiv Folia* may be treated as a separate tradition due to the objective linguistic circumstances. However, this author also includes *Kyiv Folia* into the group of OCS manuscripts [13].

The texts that are found in less than 3 types of sources are the following ones: *Samuil Inscription, Preslav Inscription, Macedonian Fragment, Martyrologium Odonis, Ostromir Gospels, Nikolje Gosples, Vatican Palimpsest, Bojana Palimpsest, Sinai Fragment, Enina Apostolos, Psalterium Demetrii Sinaitici, Missale Sinaiticum, Saint Petersburg Octoechos, Zograph Palimpsest, Codex Zographiensis—B, Vita Constantinii, Vita Methodii*, and *Prague Fragments*.

Nikolje Gospels, as stated earlier, has been included by the author into their text collection, according to non-linguistic reasons [20]. To use it, while creating a corpus, when there is no agreement, seems to be irrational. The agreement on addition of *Ostromir Gospels*, on the other hand, is strictly negative, it is widely accepted as the East Church Slavonic manuscripts. This also applies to the *Prague Fragments* that belong to Czech Church Slavonic, which is explicitly stated in the source [16].

Saint Petersburg Octoechos is a manuscript that is known mostly by East Slavic and South Slavic copies, therefore, it is preserved in a form that has been under a heavy influence of the corresponding groups of idioms. It seems that the texts it represents are rather texts in East Church Slavonic or South Church Slavonic [23]. *Octoechos*, thus, linguistically is not the OCS manuscript, and thus is not going to be included into the corpus.

Table 1 Linguistic analysis of *Samuil Inscription* [33] by the criteria from [13]. Here and then, in the article Roman script is used. Original Cyrillic examples are going to be available as part of supplementary materials

Original group	Reflex	Sample
tj*/dj*	–	–
**óRC*	*raC*	*rabŭ* ‘slave’
tl*/dl*	–	–

Addition of *Psalterium Demetrii Sinaitici* [15] and *Missale Sinaiticum* [26] also does not seem to be possible. Their texts are not published; there are only fragmentary samples of these manuscripts language. Therefore, any conclusion on the nature of this language is a baffling problem at the moment. Before the entire text is published the addition of these texts into the corpus should be postponed.

In case of the other manuscripts, additional research should be conducted, because there are neither linguistic or historical reasons nor agreement on whether to include these documents into the OCS corpus, or to exclude such a possibility. Thus, it is necessary to analyze them by the criteria, given in [13].

Table 1 clearly indicates that there is no definition for attributing the text to the OCS reflex, and the text in general is not big. By the only fixed reflex **óRC* > *raC* it can only be stated that it is either the South Slavic idiom or Czech or Slovak, but not more than that. This is why *Samuil Inscription* addition by now is unlikely.

The situation with *Preslava Inscription* [24] is even more complicated. There are no specific OCS features, or at least shared with some other Slavic languages, in this manuscript. It cannot be excluded from the list of possible additions to the corpus, but cannot be included by the same criteria. It is impossible to verify which language exactly the author used.

The language of *Macedonian Fragment*, thus, is OCS. The original groups**tj* and **dj* are reflected as *št* and *žd* in the manuscript, which is the main reason to attribute any text as OCS. Additional criteria also support the hypothesis. These are transition of **órC* > *raC*, and simplification of consonant group **tl*/**dl* (namely transition *dl* > *l*). *Macedonian Fragment*, thus, should be placed into the prepared OCS corpus (see Table 2).

The text has comparatively little amount of **tj* reflexes, and only one **dj* reflex, which makes significantly harder its linguistic attribution. Additional signs of linguistic origin of the text are hardly present as well, and only help to conclude that it is either South Slavic or Czech or Slovak. However, the existing descriptions point

Table 2 Linguistic analysis of *Macedonian Fragment* [22] by the criteria from [13]

Original group	Reflex	Sample
tj*/dj*	*št/žd*	*ašte* ‘if’, *obręštetŭ* ‘find’ *nuždi* ‘need’, *ižde* ‘and’
**óRC*	*raC*	*razumæjetŭ* ‘understand’, *razumŭ* ‘mind’, *rabŭ* ‘slave’
tl*/dl*	*l*	*molǫ* ‘pray’

Table 3 Linguistic analysis of *Martyrologium Odonis* [31] by the criteria from [13]

Original group	Reflex	Sample
**tj/*dj*	*št/ž*	*sǫšt(i)* ‘existing’, *nosęštija* ‘bearing’ *rožĭsto* ‘Christmas’
**óRC*	*raC*	*razdrušĭno* ‘in a destroying manner’
**tl/*dl*	–	–

out that the majority of *Martyrologium Odonis* inscriptions has been made by East Slavic writers in the post-OCS period. This, according to the authors, is highlighted by additional *d* in *razdrušĭno* ‘in a destroying manner’, and writing of name *kjurilŭ* ‘Cyril’ via *uk* preceded by *j* [28]. Primarily these additional features combined with the only fixated reflex **dj* > *ž* (and not *žd*) point out the East Slavic nature of the text, which, thus, should not be included in the OCS corpus (see Table 3).

Sinai Fragment text is comparatively short and is not preserved well enough, so the analysis by the criteria from [12] becomes much harder. Table 4 represents the full list of lexemes from the text that can be used for this analysis. For the first of the additional criteria, **tl/*dl* reflex, there is no material. For the second one, **óRC*, we have found only one lexeme. For the main criterion, **tj/*dj* reflex, there are three lexemes, but two of them are restored, nevertheless, the reflexes themselves, *št* and *žd*, are visible quite clearly. The words in which they were restored by scholars, were not used. However, by the preserved features the text certainly belongs to OCS (see Table 4).

The linguistic features of *Enina Apostolos* (see Table 5) point out its OCS origin. Primarily these are **tj* and **dj* reflexes, *št* and *žd* respectively. Additional features, **óRC* и **tl/*dl*, also indicate that this is the South Slavic idiom (excluding one by one its attribution to the East Slavic and West Slavic groups). Consistent with all of these criteria, *Enina Apostolos* is an OCS text and will be included into the corpus.

Table 4 Linguistic analysis of *Sinai Fragment* [10] by the criteria from [13]

Original group	Reflex	Sample
**tj/*dj*	*št/žd*	*poštędi* ‘spare’, *[svæ]štǫ* ‘candle’ *(d)[a]ždĭ* ‘give’
**óRC*	*raC*	*rabŭ* ‘slave’
**tl/*dl*	–	–

Table 5 Linguistic analysis of *Enina Apostolos* [27] by the criteria from [13]

Original group	Reflex	Sample
**tj/*dj*	*št/žd*	*nemoštnu* ‘infirm’, *mogǫštomu* ‘able’, *noštĭ* ‘night’ *præžde* ‘befire’, *tuždego* ‘alien’, *odeždǫ* ‘clothes’
**óRC*	*raC*	*razumŭ* ‘mind’, *raba* ‘slave’, *rabotǫ* ‘work’
**tl/*dl*	*l*	*molitvy* ‘prayer’

Table 6 Linguistic analysis of *Zograph Palimpsest* [21] by the criteria from [13]

Original group	Reflex	Sample
tj*/dj*	*št*	*cædališta* ‘that thing what somebody sits on’, *prodajǫštixŭ* ‘selling’, *ašte* ‘if’
**óRC*	*raC*	*razboinikomŭ* ‘bandits’, *rabŭ* ‘slave’
tl*/dl*	*l*	*molitvæ* ‘prayer’

The next texts, *Zograph Palimpsest* and *Codex Zographiensis* are very similar because *Codex Zographiensis* is written over *Zograph Palimpsest* with Cyrillic script. However, these are two different texts. *Codex Zographiensis—B* is much bigger (which is seen, for instance, by the diversity of reflexes that have been analyzed), and in coinciding parts different fragments are skipped. This is why *Zograph Palimpsest* and *Codex Zographiensis* should be analyzed separately.

The tricky part of *Zograph Palimpsest* analysis is a total lack of words, by which one may check, how **dj* reflected in the language of the manuscript. However, high frequency of *št* < **tj* reflex, and additional features, **óRC* and **tl*/**dl* reflexes, in general demonstrate that by the defining criteria of dental consonant with jot groups reflexes *Zograph Palimpsest* linguistically is an OCS text, therefore, its addition to the corpus is going to make a positive contribution to the representativeness of the latter (see Table 6).

The set of reflexes that is present in *Codex Zographiensis—B* is quite large and allows to infer that the text surely belongs to the OCS canon. It is of special importance that reflexes **tj* > *št/* and **dj* > *žd* allow to make a conclusion about the language of the manuscript. On this basis *Codex Zographiensis—B* is to be included into the OCS corpus (see Table 7).

It can be seen that judging by additional features the text may be considered OCS, or, at the very least, Church Slavonic. However, there are East Slavic reflexes **dj* > *ž* and **tj* > *č*, cf. *čjužei* < *čjužĭ* ‘alien’ with OCS *štuždĭ* ‘alien’. The reflex **dj* > **žd* is not encountered in the text at all, so, *Vita Methodii* in the known copy is not an OCS text (see Table 8).

Vita Constantinii is a manuscript that contains significantly less East Slavic features than Vita Methodii (see Table 9). **tj* and **dj* groups reflexes are typical for the East South Slavic area, which is supported by the additional features **óRC* and **tl/*dl*. The text does not indicate cases, when etymologically restored group [s’č’] is presented by grapheme *št*, as well as lexemes that contain non-South Slavic

Table 7 Linguistic analysis of *Codex Zographiensis—B* [6] by the criteria from [13]

Original group	Reflex	Sample
tj*/dj*	*št/žd*	*xoštetŭ* ‘wants’, *otvæštavŭ* ‘answered’, *obrěšteši* ‘having found’ *tuždixŭ* ‘alien’, *præžde* ‘before’, *daždĭ* ‘give’
**óRC*	*raC*	*razdræšite* ‘allow’, *rabe* ‘slave’
tl*/dl*	*l*	*molæše* ‘praying’, *pomolitŭ sę* ‘have a pray’, *molitvæ* ‘prayer’

Table 8 Linguistic analysis of *Vita Methodii* [6] by the criteria from [13]

Original group	Reflex	Sample
tj*/dj*	*št*, *č/ ž*	*ašte* ‘if’, *nošti* ‘night’*, pomoštĭ* ‘help’ *čjužei* ‘alien*’* *režajetŭmĭ* ‘gives birth’, *præže* ‘before’, *mežju* ‘between’*, dažĭ* ‘give’, *žaži* ‘thirst’
**óRC /* **óLC*	*raC, laC, alC*	*razumæti (*and other words with prefix *raz*- ‘rahz-’ < **orz*) ‘understand’, *ravnŭ* ‘equal’, *rabi* ‘slaves’ *lakomĭstva* ‘delicasies’ *alŭkani* ‘greed’
tl*/dl*	*l*	*molitvū* ‘prayer’ (with all its grammatical forms)

Table 9 Linguistic analysis of *Vita Constantinii* [6] by the criteria from [13]

Original group	*Reflex*	Sample
tj*/dj*	*št* (in form of single letter) */ žd*	*xoštetĭ* ‘want’, *ašte* ‘if’, *iskušajušte* ‘seducing’, *pomoštiu* ‘help’, *otvæštaše* ‘answering’, *pasuštii* ‘shepherding’ *tuždĭ* ‘alien’, *daždĭ* ‘give’, *ugaždaje* ‘pleases’, *viždu* ‘see’, *meždu* ‘between’, *graždanŭ* ‘citizens’
**óRC*	*raC, laC*	*razumĭ* ‘mind’, *razluči* ‘separate’, *rabŭ* ‘slave’, *razorenu* ‘devastated’, *različii* ‘difference’ *lakomja* ‘tasty’
**tl/*dl*	*l*	*moljaše* ‘praying’, *molitvaxŭ* ‘prayers’, *moliti* ‘pray’

reflexes of Old Slavic groups with liquid consonants. The text may be called OCS by the criteria from [13].

However, its language exhibits another feature that casts doubts on the correctness of the hypothesis. It is graphemes that are seen in the place of etymologically restored nasal vowels. Where one expects big jus, there is *u*. Where one expects jot-preceded big jus, there is *ju*. Where one expects small jus, or jot-preceded small jus, there is *e* or *je* (inconsistently, cf. *jezūkŭ—ezūkŭ* ‘language, people’). The only jus that appears in the context *vŭprošĭšem’ žę nækoimĭ* ‘asked somebody’, where, as it seems, indicates not the nasal vowel, but its reflex, being encountered throughout the whole text.

It seems that the text is a very late copy of the OCS original [11], made in South Slavic area (the latter is highlighted by the forward nasal reflex). Its linguistic features do not allow its inclusion into the corpus.

It is impossible to analyze *Vatican Palimpsest* and *Bojana Palimpsest* at the time of writing this article for various reasons including the epidemiological situation and the situation around Ukraine.

4 The OCS Corpus Creation: The Building of System

The concept of a system that represents the corpus in a digitalized form with short characteristics of its main parts has been presented in [1]. During the creation of the corpus of OCS, only a limited set of them has been used, namely field creation submodule, annotating text submodule, text features visualization submodule, and search submodule. This system was chosen against the more popular and robust ones, such as Sketch Engine, as it is more flexible to the needs of the researcher of low-resource ancient languages [2]

Field creation submodule is currently being edited, however, its functionality already includes field creation, which includes giving description, choosing, whether it is a field for a manuscript, its part, the segment of the latter, token, or grapheme. It is possible to connect fields to values of other fields, which may be used for manual tagging. Fields may have a user-given set of values (cf. field of creation period), or arbitrary value, given to each unit of database during the process of tagging (cf. lemma field). The fields may be assigned one time for one tagging (as part-of-speech field) or more than once for one tagging (as a lexical group field). Single fields of one level may be interconnected (for instance, a token field with a token field).

Text annotation submodule depicts the chosen by the user manuscript part, split via HTML into segments, that are separated into tokens, which themselves are separated into graphemes. The segment, the token and the grapheme each may be tagged. Identical tokens may be tagged all at the same time, through the first encountered. Moreover, any token may be tagged in different ways in case it is impossible to select the exact value in category.

Tagging of the segment is done after pressing the button. Tagging of the segment is done after double-clicking of a token with left mouse button, and the grapheme is tagged when the right mouse button is pressed. In any case the user may tag the unit only with the fields that are designed for it. In addition, if the chosen value of a chosen unit is connected to other fields, they appear lower, so one may tag a unit with them as well.

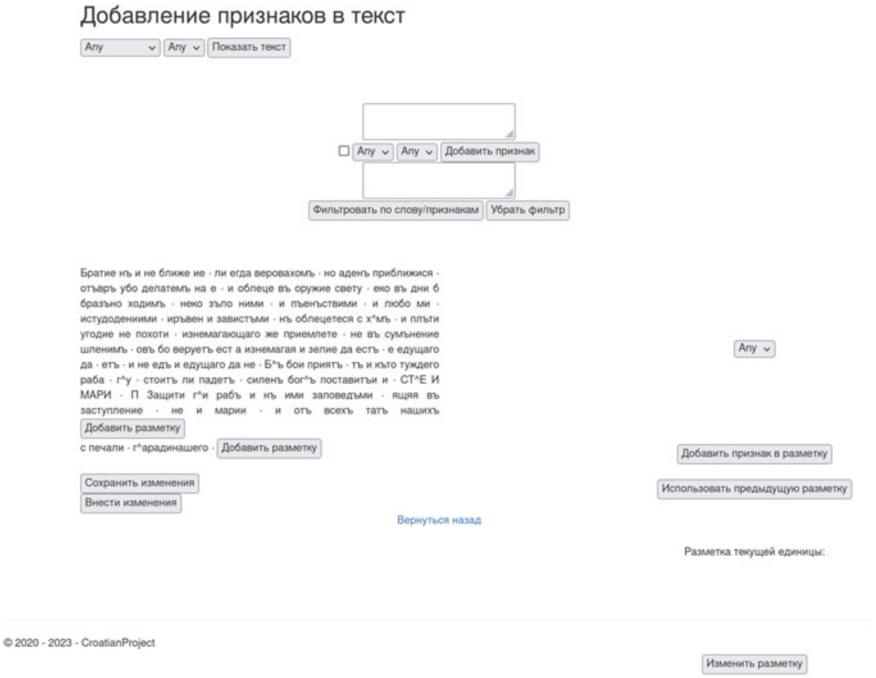
During the tagging process special attention should be paid to the units of some classes, or the units that contain a certain grapheme sequence. For the purpose of filtration of existing or not existing features, or character sequences, chosen by the user, there is a module in the upper part of the page.

Text annotation submodule window is presented in Fig. 1.

Text features visualization submodule provides the user with the possibility to depict a created annotation of a chosen manuscript part by segments, tokens, and graphemes. The function of filtering the tokens by sequences of characters within them is available for this part, as well as for text annotation submodule.

By now, the only way to visualize text features is to highlight with color the units that comply with the criteria, provided by the user, namely, having or not some features. The example representing the highlighting of numbers is given below:

*VŬ **12** DĬNĬ KLIMENTA ~Bŭ iže nū ..*


Добавление признаков в текст
Any
Any
Показать текст
□ Any
Any
Добавить признак
Фильтровать по слову/признакам
Убрать фильтр
Any
Добавить разметку
Добавить разметку
Сохранить изменения
Внести изменения
Вернуться назад
Добавить признак в разметку
Использовать предыдущую разметку
Разметка текущей единицы:
© 2020 - 2023 - CroatianProject
Изменить разметку

Братие нъ и не ближе ие · ли егда веровахомъ · но адень приближися · отъврь убо делатемь на е · и облеце въ оружие свету · еко въ дни б бразъно ходимъ · неко зъло ними · и пъенъствими · и любо ми · истудодениими · иръвен и завистъми · нъ облецетеся с х^мъ · и плъти угодие не похоти · изнемагающаго же приемлете · не въ сумънение шленимъ · овъ бо веруетъ ест а изнемагая и зелие да естъ · е едущаго да · етъ · и не едъ и едущаго да не · Б^ъ бои приятъ · тъ и къто туждего раба · г^у · стоитъ ли падетъ · силенъ бог^ъ поставитъи и · СТ^Е И МАРИ · П Защити г^и рабъ и нъ ими заповедъми · ящяя въ заступление · не и марии · и отъ всехъ татъ нашихъ

с печали · г^арадинашего ·

Fig. 1 Text annotation submodule window

The main part of the corpus is the search submodule. On this stage it is possible to search by separate manuscripts: creating the subcorpora is the task for the future versions of the system. Apart from that, user may search by token parts, as well as by preliminary made manuscript annotation. It is allowed to use annotation of manuscripts, their parts, segments, tokens, and graphemes.

The output includes all the tokens that match the given parameters. The pagination is not yet included into the submodule, all the contexts are presented on the single page. Every token is given in *KWIC (Key Word In Context)* format. The first string of each output unit may find the manuscript ID, its name and part. The next strings are taken by segment tokens before the key word. After that the separate string is taken by the key word. Finally, the rest of the segment is given.

The example of unit output is provided below. It is a fragment of a query for giving all the lemmata in the corpus that start with ~.

7. 4_ZogrB_tagged: 4_ZogrB_tagged išĭdŭ že rabŭ tŭ obræte jedinogo otŭ klevrætŭ svoixŭ iže bæ dlŭžnŭ jemu

~100~

pænęzŭ.

5 Advantages of Possible Corpus Applying in the University OCS Learning and Studies

Creation of the corpus is not the task which can be solved at once and in a short time, but even at the initial stage, available in the net and supplied with systems for search and excerpts text fragment it might provide significant support to OCS teachers, of whom there are hundreds in Russia and supposedly thousands around the world. With the first texts in database, it is possible to use the corpus for teaching purposes, providing students with the materials and the tool for demonstrating the features they are studying. This can lead to creating the corpus-based textbooks when more texts are digitalized.

The corpus can be addressed to students not via teachers but directly. Considering that thousands of students start learning OCS every year, it is rare that such studies turn into independent students' research. Among numerous reasons for that we should mention difficulties in access to OSC texts with the length adequate for research, not just for achieving elementary competence to process OSC texts. The texts included into textbooks are selected to demonstrate some language features or specificity of a particular monument (cf. fragments from OCS texts in the most voluminous textbooks such as [18, 28, 34]), while professional publications of OCS monuments are usually published in limited editions and not every institution where OCS is taught have them at its disposal.. Besides, when a researcher works with traditional publications it takes them a lot of time to discover the text pieces, they are interested in. If OCS texts are assembled in the corpus, it eliminates these problems making a very wide textbase available and supplying it with search and excerpts text fragment systems. Existing of the corpus might let even bachelor students solve some research tasks in their term papers. These could be various confrontations of facts occurring in different parts of the same monument or in different monuments, e. g. comparison of spelling, of grammar forms in use, of fluctuations in inflection, differences of word usage in translations of the same source text etc. The research also becomes much easier with the corpus, whether this is a test of an existing model against the real data, the opportunity more suitable for the generative linguistics, or formulating the theory from the large set of data, the possibility that is negated by the generative linguistics, but still seeming to be valid, as shown by corpus linguists [8]. The first articles of the students may present their experience with the corpus, either with the participation in its creation, or with the analysis of already existing parts of the corpora.

Otherwise, since the corpus is being created for a longer period, also after it becomes available for researchers it is still being improved and topped up. Therefore, the work on its improvement and completion may become a good practice for the students—at least at the universities providing technical base for its creation. Modern network connections make this work available for every person or institution which manages to join the appropriate network. It would enable students not only use this corpus but also help with that. Each university, in fact, each teacher may fork the corpus and add the texts appropriate for the preferences and goals. The network between these corpora may also be organized for sharing knowledge and tagging exchange.

The created corpus makes possible the transition from the first experiments to the systematic formation of the corpus that is available for students and scholars. It is going to be a learner corpus that, according to some scholars, may also be used for other purposes, for instance, for training of OCR neural networks [3]. For this, the experience of African researchers may be very useful, because corpus linguistics in the Sub-Saharan space is highly specialized in learner corpus creation [5].

This is not the full range of uses of the corpus in university practice, yet, these seem to be the most obvious and probably popular ones. Other usages may be found through discussion and actual implementation of the corpus into actual work process. It is hard to believe that revitalization, as with Nasa Yuwe [14], should happen. Yet the question of developing the tools, enabling students to participate in more and more complicated activities [14], may be of big importance for the future of the OCS corpus.

6 Conclusion

The OCS corpus creation has been started, and the web interface for the interaction is given. The program is ready for the new modules incorporation. For instance, each text part may be accompanied by the image of its original presented via a separate page or a pop-up.

The whole OCS canon was analyzed, and the texts were ordered for the addition to the corpus. The draft for the corpus split into subcorpora is provided, as the OCS canon is, as it was discovered during this article preparation, quite a heterogeneous list of texts. The further investigation into this matter is going to be the material for the following research.

The possible applications for the currently created OCS corpus are wide. It may be used for the corpus-based research, including investigation into the issue of heterogeneity of the OCS canon. Students may learn it, tag it, or perform their own projects. This probably may be the beginning of a new time for the OCS research, at the very least, we hope so.

References

1. Afanasev, I. (2020). Corpus-dictionary system: Introducing a concept. In A. A. Kibrik, V. Gusev, D. Zalmanov (Eds.), *International conference "Linguistic Forum 2020: Language and artificial intelligence, 12–14 November 2020* (pp. 10–11). Institute of Linguistics RAS, Moscow, Russia.
2. Afanasev, I. (2020). A corpus-based approach in archaeolinguistics. *Journal of Applied Linguistics and Lexicography, 2*(2), 147–159.
3. Alfaifi, A., & Atwell, E. (2013). Potential uses of the Arabic learner corpus. In Leeds language, linguistics and translation PGR conference (pp. 1–3). University of Leeds.

4. Bally, C., & Sechehaye, A. (Eds.). (1959). Course in general linguistics (Trans. W. Baskin). The Philosophical Society.
5. Carstens, A., & Eiselen, A. (2019). Designing a South African multilangual learner corpus of academic texts (SAMuLCAT). *Language Matters*, *50*(1), 64–83.
6. CCMH–Corpus Cyrillo-Methodianum Helisingiense. Retrieved 19 May 2022, from https://www.helsinki.fi/slaavilaiset/ccmh/index.html
7. Cyrillomethodiana–Bulgarian Historical Corpus. Retrieved 29 Sep. 2022, from https://histdict.uni-sofia.bg/
8. Divjak, D., Sharoff, S., & Erjavec, T. (2017). Slavic corpus and computational linguistics. *Journal of Slavic Linguistics, 25*(2), 171–198
9. Freeman, S., Herron, J. C.: Evolutionary Analysis. 3rd edn. Pearson Education, Upper Saddle River, NJ (2004).
10. Glibetić, N.A. (2015). New eleventh-century Glagolitic fragment from St Catherine's monastery: The midnight prayer of early Slavic monks in the Sinai. *Археографски прилози, 37*, 11–47.
11. Grivec, F., & Tomšić, F. (1960). Constantinus et Methodius Thessalonicenses, Fontes. *Radovi Staroslavenskog instituta, 4*(4), 13–276.
12. Groenewald, H. (2007). *Automatic lemmatisation for Afrikaans*. North-West University.
13. Kamphuis, J. (2020). Verbal aspect in old Church Slavonic. Brill.
14. Martinez, L. M. S., Cobos, A. C., Muños, J. C. C., Curieux, T. R., Herrera-Viedma, E., & Peluffo-Ordoňez, D. H. L. (2018). Building a Nasa Yuwe language corpus and tagging with a metaheuristic approach. *Computación y Sistemas, 4*(3), 881–894.
15. Tarnanidis, I. (1988). The psalter of Dimitri the Oltarnik. In The Slavonic manuscripts discovered in 1975 at St. Catherine's monastery on mount Sinai (pp. 91–100). Hellenica Association for Slavic Studies
16. TITUS–Thesaurus Indogermanisher Text- und Sprachmaterielen. Retrieved 29 Sep. 2022, from http://titus.uni-frankfurt.de/indexe.htm
17. TOROT. Retrieved 19 May 2022, from https://nestor.uit.no/.
18. Weingart, M., & Kurz, J. (Eds.). (1949). Texty ke studiu jazyka a písemnictví staroslověnského (2nd ed., revised). Slovanský Seminar Karlovy University
19. Baranov, V. A. (2019). K voprosu ob ispol'zovanii statisticheskikh metodov dlya poiska kollokatsy i kolligatsy v drevneyshikh slavyanskikh tekstakh (na materiale glagoli-cheskikh rukopisey korpusa «Manuskript»). In *SLOVO 69* (pp. 1–33).
20. Vayan, A. (1952). Rukovodstvo po staroslavyanskomu yazyku. Izdatel'stvo inostrannoy literatury, Moscow.
21. Dobrev, I. (1971). Palimpsestovite chasti na Zografskoto Evangelie. In: Konstantin-Kiril Filosof. Dokladi ot simpoziuma, posveten na 1100-godishninata ot smrtta mu (pp. 157–164). Izd-vo na Blgarskata akad. na naukite, Sofia.
22. Il'insky, G. A. (1906). Makedonsky listok. Imperatorskaya Akademiya Nauk, Saint Peters-burg.
23. Krasheninnikova, O. A. (2006). Drevneslavyansky Oktoikh sv. Klimenta, arkhiepiskopa Okhridskogo: Po drevnerusskim i yuzhnoslavyanskim spiskam XIII—XV vekov. Yazyki slavyanskikh kul'tur, Moscow.
24. Medyntseva, A. A., & Popkonstantinov, K. (1985). Nadpisi iz Krugloy tserkvi v Preslave. Bolgarskaya akademiya nauk, Sofia.
25. Meye, A. (1934). Obshcheslavyansky yazyk. Izdatel'skaya gruppa «Progress», Moscow
26. Miklas, K., Sadovski, V., Khyurner, D., & Vandl, F. (2022) Sinaysky glagolichesky Litur-giary («Missal»). Retrieved 29 Sep. 2022, fromhttp://manuscripts.ru/mns/portal.main?p1=67
27. Mirchev, K., & Kodov, K. (1965). Eninski apostol: Staroblgarski pametnik ot XI v. Izd-vo na Blg. akad. na naukite, Sofia.
28. Nikoliħ S. (2001). Staroslovenski jezik: II: Primeri sa rechnikom (13th ed izd). Trebnik, Belgrade.
29. Polivanova, A. K. (2013). *Staroslavyansky yazyk: Grammatika: Slovari*. Universitet Dmitriya Pozharskogo.

30. Selishchev, A. M. (1951). Staroslavyansky yazyk: Chast' pervaya: Vvedenie: Fonetika. Gosudarstvennoe uchebno-pedagogicheskoe izdatel'stvo ministerstva prosveshcheniya RSFSR, Moscow.
31. Sobolevsky, A. I. (1910). Materialy i issledovaniya v oblasti slavyanskoy filologii i arkheologii. Imperatorskaya Akademiya Nauk, Saint Petersburg.
32. Tseytlin, R. M., Vecherka, R., & Blagova, È. (Eds.). (1994). Staroslavyansky slovar': (po ruko-pisyam X–XI vv.). Ok. 10000 sl. Rus. yaz., Moscow.
33. Uspensky, F. I. (1899). Nadpis' tsarya Samuila. In Izvѣstiya Russkago Arkheologicheskago instituta v Konstantinopolѣ 4 (pp. 1–4).
34. Shulezhkova, S. G. (2013). Khrestomatiya po staroslavyanskomu yazyku: Teksty, slovar', fonoprilozhenie (3rd ed.). Flinta, Nauka.

Core Coordination Units in Macro- and Microdiachrony: Experimental Data

Yuliya Auseichyk

Abstract Based on the Frantext corpus data, the article compares relative frequency indicators for traditionally distinguished French coordination conjunctions, which are qualified in the study as core coordination units. It is shown that on a macroscale with a hundred-year interval (1100 to present) and on a microscale with a ten-year interval (1800 to present) the graphs of relative frequency curves for these units reflect respectively the result and the process of how nomination of various relations between extralinguistic entities was cognitive mastered. The "attenuation effect" of the frequency of these units over the centuries-old language history is demonstrated. Individual trajectories in the use of core coordination units are described: an increase and decrease in their frequency with different intensity in macro- and microdiachrony. We reveal the leading role of the additive *et* in forming the coordinative relation before 1800, and an increase in the use of the adversative *mais* over the last decade. The observed general trend to reduce the frequency of core coordination units occurs due to the manifestation of multidirectional individual trends in designating the essence of additive, adversative, disjunctive and causal relations.

Keywords Coordination · Core units · Corpus data · Usage trends · Relative frequency · Diachrony

1 Introduction

Speakers' desire to explicitly express logical connections is predetermined by continuous development mental activity of native speakers, which eventually deepens ideas about the surrounding reality, making them more complicated. In everyday life, human cognition moves from reflecting observed relations to conveying abstract, speculative relations [1], which predetermines the use of units existing in the language system to denote both observed relations between entities and speculative, increasingly complex relations between situations in the extralinguistic reality [2–4]. These

Y. Auseichyk (✉)
Minsk State Linguistic University, Minsk 220034, Republic of Belarus
e-mail: ovsei77@rambler.ru

P. Eismont et al. (eds.), *Literature, Language and Computing*,
https://doi.org/10.1007/978-981-99-3604-5_5

changes, «grounded in cognitive processes and usage factors» [5: 9], take place very slowly, over several centuries. Thus, the system of French coordination conjunctions, which traditionally includes units *et* 'and', *ou* 'or', *mais* 'but', *ni* 'neither', *donc* 'so', *or* 'thus', *car* 'because', was formed gradually and unevenly in the early period of the language development (9th to sixteenth century) and persists to the present.

We believe that significant changes have taken place in the system of coordination units in the course of historical language development, which can be detected by establishing their quantitative characteristics. By quantitative characteristics we mean the usage of these units, i. e. their frequency.

2 Literature Review

Seven connecting units (Fr. *outils de liaison*) *et, ou, mais, ni, donc, or, car*, have been functioning as linkers since the 9th or tenth century. These units go back to the Latin correlates (*et < e/et, ni < ne/ni < nec, ou < o < aut, car < quare, or < Hac hora, mais < magis, donc < dum* [6]) and are inherited from Latin (in the terminology of F. Bruno «*héréditaires*» [7: 716]). We qualify these conjunctions as core coordination units. Conjunction *et* denotes addition (in negative contexts, conjunction *ni* acts as a functional counterpart of additive *et*); the conjunction *ou* expresses disjunctive relations; the conjunction *mais* shows adversative relations. Three conjunctions *car, donc, or* indicate causality [8, 9].

In modern Roman studies, starting with the fundamental work on coordination by G. Antoine [10], functionality of coordinators has been the objective of numerous research papers, see e.g. [11]. Individual coordinators were studied from the diachronic standpoint, for more details see [12: 944–963]. Meanwhile, comparative analysis of core coordination units in terms of their usage throughout the centuries-old history of the French language has never been the subject of scientific discussion.

Development of corpus linguistics has promoted interest towards the processes of diachronic changes in languages, with representative selection of contexts making it possible to "grasp" some patterns of system evolution (for details regarding general types of corpus analysis see [13], as well as [14–16]).

The present corpus-based research implies quantitative analysis of empirical material, first and foremost. Interpretation and comparison of quantitative frequency-related indicators demonstrated by the studied units in diachronic perspective allow us to gain new data how the system of French core coordination units has evolved.

3 Methodology

The paper studies frequency characteristics of French core coordination units *et, ou, mais, ni, donc, or, car*.

Our analysis relies on data obtained from the Frantext French National Corpus [17], which includes 5,555 texts of different genres of the 9th through twenty-first century with a total volume of 264 million word forms. The corpus texts are structured according to their creation date:

- years 1100–1199—39 documents;
- years 1200–1299—35 documents;
- years 1300–1399—118 documents;
- years 1400–1499—162 documents;
- years 1500–1599—187 documents;
- years 1600–1699—636 documents;
- years 1700–1799—702 documents;
- years 1800–1899—1152 documents;
- years 1900–1999—2222 documents;
- years 2000– curr. time—301 documents

Considering that its digitized documents cover a thousand-year history of language development from the ninth century to the present, we proceed from the fact that the corpus can be deemed as representative to reveal diachronic changes in the system of core coordination units. We would like to point out that this paper analyzes written documents, without taking into account vernacular language.

Considering the uneven representation of documents across different subsections of the Frantext Corpus and in order to optimize research procedures, we limited the corpus material to fragments represented by one million word forms (taking for granted that a selection of one million word forms is representative enough for units as frequent as prepositions and conjunctions, see [18, 19]).

We assume that for every core coordination unit individual relative frequencies reflect multidirectional individual trends in their usage, manifested in a regular decrease and increase in the frequency of units with varying intensity throughout different historical periods of language development, thus establishing a general trend.

To confirm our hypothesis, we conducted the experiment to compare relative usage frequencies for the selected units in macrodiachrony (in a diachronic perspective, ranging from 1100 to the present at a hundred-year's interval) and in microdiachrony (within a two-hundred-year synchronous perspective, ranging from 1800 to the present at a ten-year's interval). To establish relative frequencies for *et, ou, mais, ni, donc, or, car*, a smoothed-shape curve was used. The results of the experiment are presented below.

4 Frequency Characteristics of the Core Coordination Units in Macrodiachrony

The graph showing relative frequency curves for *et, ou, mais, ni, donc, or, car* from 1100 to the present measured at a century's interval represent their frequency distribution in macrodiachrony as shown on the diagram (see Fig. 1).

It is evident from the diagram in Fig. 1 that over the historical development of the French language, the core coordination units are gradually decreasing in frequency. From 1100 through 1800, unit *et* plays a leading role in rendering coordinative

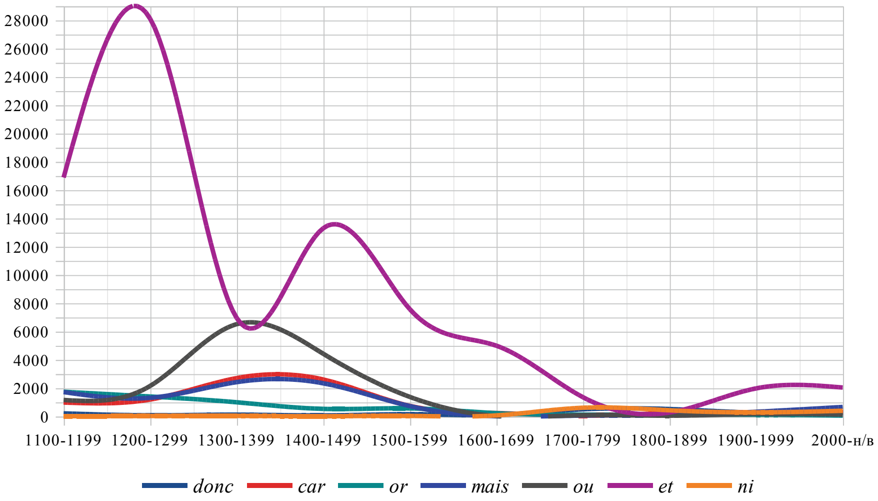
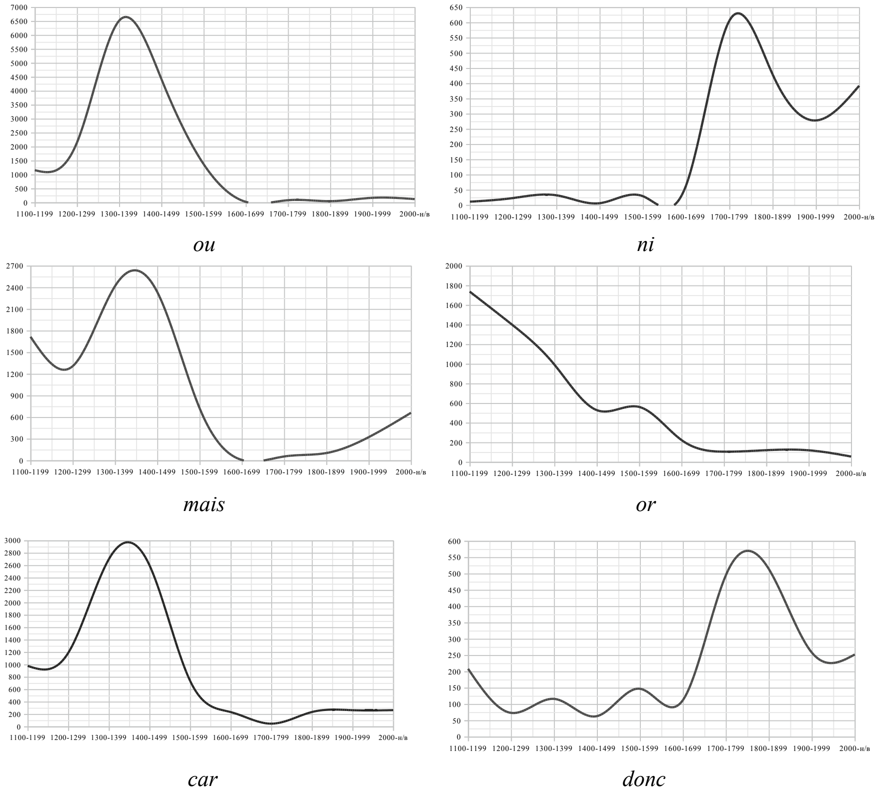


Fig. 1 Relative frequency of core coordination units in macrodiachrony (1100–present)

relations: its relative frequency significantly exceeds the frequency of other units. At the same time, the search recorded both a sharp increase in the frequency of *et* from 1200 through 1299 followed by a sharp decrease from 1700 through 1799. (cf.: 28,106 vs. 1,349 occurrences per million word forms, respectively). Since 1800, the Modern French period shows some gradual flattening of all frequency curves, arising from minimal differences in quantitative indicators for the highest frequency conjunction *et* and the lowest frequency conjunction *ni* (cf.: 270 vs. 429 occurrences per million word forms in 1800 through1899). Thus, we observe the «frequency attenuation» effect (the term coined by Baranov [20: 156]). In other words, *ou*, *mais*, *ni*, *donc*, *or*, *car* yield more flattened curves because the number of occurrences for *et* is so large, identical scaling results in significant flattening of the distribution curves for less frequent units. Another example is the thousandfold difference between the most frequent conjunction *et* and the least frequent conjunction *ni* in 1200 through 1299. (cf.: 28,106 vs. 28 occurrences per one million word forms, respectively).

The unit frequency curves shown in the diagram in Fig. 1 illustrate the general usage dynamics and do not allow us to trace frequency trajectories for every individual unit. In this regard, it seems appropriate to separately consider frequency distributions for six less common core units in macrodiachrony and how those correlate with maximum and minimum relative frequencies (see diagrams in Fig. 2).

The diagram Fig. 2 shows individual usage trajectories for *ou*, *mais*, *ni*, *donc*, *or*, *car* demonstrate changes in their relative frequency. Despite apparent differences (sharp fluctuations in the frequency of some units and more moderate fluctuations in the frequency of others for different periods, a shap increase or a flat start, or a sharp drop, etc.), the constructed graphs have a number of similarities (convex and concave curves for certain time intervals). Types of the curves reflect the nature of frequency changes: convex curves correspond to a gradual increase in frequency,



ou *ni*

mais *or*

car *donc*

Fig. 2 Individual trajectories of the core coordination units in macrodiachrony (1100–present)

while concave curves correspond to more dynamic and sharp changes according to A. Baranov [20: 148]. The analysis shows that during the early stage of language development, five of the seven core units *et*, *ou*, *mais*, *or*, *car* manifest a surge in their usage, while the usage growth of *ni* and *donc* occurs at the beginning and the middle of the eighteenth century.

The curves for *ou*, *car* and *mais* show "peaking" usage within the same time interval, from 1300 to 1400. Such increase in usage is explained by semantic expansion. We will refer to a few examples.

Semantic expansion of *ou* is driven by the differentiation between exclusive disjunction, e. g. *foible ou fort* 'weak or strong',[1] *yver ou esté* 'winter or summer' and inclusive disjunction, e. g. *et pour ceus qui cause ont ou auront* 'and for those who have or will have a reason' (examples hereinafter are taken from [17], unless indicated otherwise) (for more details see [21]).

Adversative unit *mais* begins to be actively used to connect two components of either opposing meaning—e. g. *Vallet ne seray plus, mais maistre* 'I will no longer

[1] Hereinafter translated by author.

be a servant, but a master', or a modified one—e. g. *Les simples prestres ne sont mie seulement entechiés, mais arcevesques, evesques et prelas*… 'Ordinary clergy are not only faithful to the faith, but they are also archbishops, bishops and prelates…' (for more details see [22]).

Semantic ambiguity of *car* is revealed when denoting consequence, e. g.: *Seignurs barons, dist li emperere Carles, / Veez les porz e les destreiz passages! / Kar me jugez ki ert en la rereguarde* 'Karl says: Oh, gentlemen-vassals! / You see the gorges and dips. / [Car] Appoint me a leader for the rearguard' (La Chanson de Roland); or reason, e. g.: *Et lors la dite Marie revint ele trouva l'enfant gueri. Car il metoit sa main destre a sa bouche et a sa teste*… 'And when Mary returned, she found that the child had recovered. For he (the healer) put his right hand to the mouth and to the head of the child…'; or justification, e. g.: *Qu'elle soit de bonne heure née! Car je sçay bien il vault miex estre De bonne heure*… 'Let her be born at dawn! For I know well that it is better to be born early in the morning…'.

In the modern language (1800 through present) we have fixed a smoothing flattening of the curve in case of *ou* and *car*, while the frequency of *mais* increases.

The established frequency distribution of core coordination units at a one hundred year's interval shows that from 1500 to 1600 the frequency of *et*, *ou*, *or*, *car* decreases, followed by subsequent balancing in modern language (1800 to present). Frequency curves for *mais*, *ni*, *donc* show differences as the number of occurrences increases after 1700.

We believe that in macrodiachrony the balancing of relative frequency curves for the observed units over long time intervals does not mean that all units are equally used after year 1800. The above mentioned results compel us to establish synchronous frequencies for the core coordination units within the microscale two-hundred year's timerange from 1800 to the present.

5 Frequency Characteristics of the Core Coordination Units in Microdiachrony

To clarify individual frequency trajectories for each unit, frequency graphs were produced for the two-hundred year's time range from 1800 to the present demonstrating frequency attenuation. For a "zoom-in" effect, frequency distribution is presented at a ten year's interval (see Fig. 3).

The presented frequency distribution of seven core units over modern French period demonstrates that frequency attenuation at the preset ten year's interval is relevant for *or*, *car*, *ni*, *donc*. Their frequency curves are characterized by smooth shape and absence of sharp frequency fluctuations. Meanwhile, additive *et*, disjunctive *ou,* and adversative *mais* do not show this trend. Their frequency varies considerably, cf.: from 17 to 2542 occurrences per million words for *et,* from 16 to 2307 for *ou*, and from 16 to 3699 for *mais.*

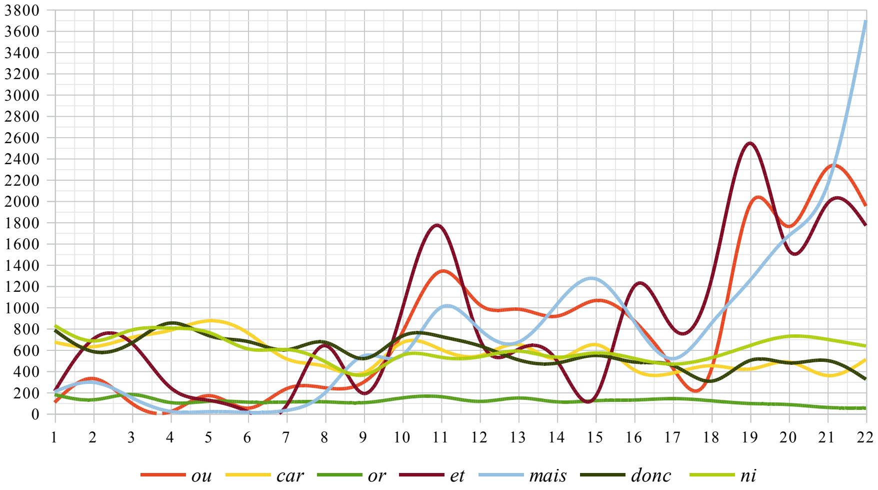


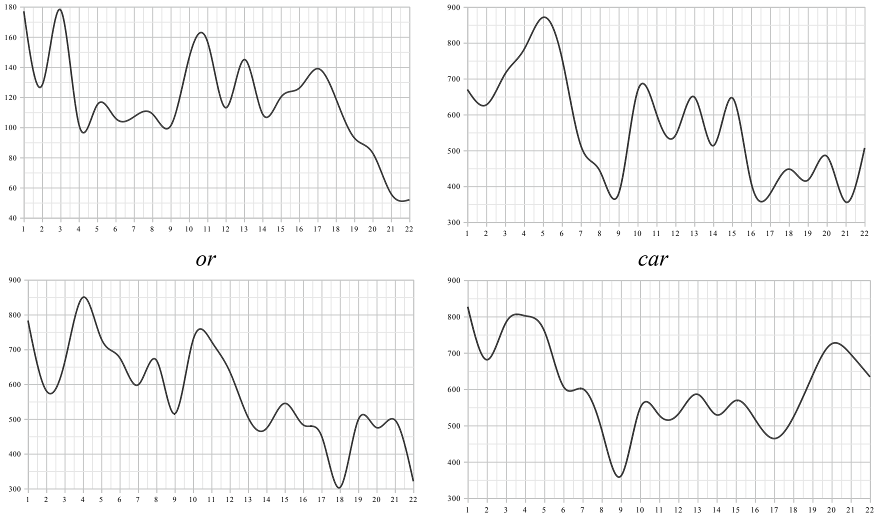
Fig. 3 Relative frequency of core coordination units in microdiachrony (1800-present)

Noteworthy is the similarity of the curve shapes for *et* and *ou*, which in fact demonstrates coincidence of convex and concave configurations and differences in frequency fluctuations after year 1960. At the same time, frequency for *mais* reaches its maximum over the last decade. At the same time, "peaking" use of *mais* is outperforms "peaks" for *et* and *ou*, cf.: 3,699 occurrences for *mais* from 2010 to present; 2,307 occurrences for *ou* in the period from 2000 to 2009 and 2,542 occurrences for *et* per one million words between 1980 and 1989. Compared to *et* and *ou*, high frequency of *mais* (3,699 vs. 1,766 vs. 1,950 occurrencies per one million word forms, respectively) over the last decade indicates the growing need among native speakers to nominate adversative relations.

As is the case for frequency distribution of core units in macrodiachrony, smooth curves for *ni*, *donc*, *or*, *car* can be explained in a similar way, namely that due to the disproportionate maximum and minimum number of occurrences. Therefore, the graphs are built taking in account maximum to minimum relative frequency ratios individually (see the diagrams in Fig. 4).

Frequency fluctuation ranges for *ni*, *donc*, *car* are comparable and span between 303 (the minimum number of occurrences for *donc* per one million word forms from1970 to1979) to 871 (the maximum number of occurrences for *car* per one million word forms from 1840 to 1849). At the same time, we can observe six "peaks" for *donc* and seven for *car*, the last occuring in the last decade. Relative frequency fluctuations for *or* are not that significant, ranging from 52 to 177 occurrences. Over the last decades *or* shows has a clear trend to decrease in frequency.

We should specifically note that the results of our experiment indicate that on the microscale units *et*, *ou*, *mais* with a higher frequency show significant frequency fluctuations with the average range of difference between 150 to 200 times; in contrast, less frequent units *ni*, *donc*, *or*, *car* manifest a very low degree of fluctuation, with a three times difference at its utmost.



or *car*

donc *ni*

Fig. 4 Individual frequency trajectories of the core coordination units in microdiachrony (1800-present)

6 Discussion

By comparing frequency fluctuations for core coordination units on a macro- and microscale, we suggest speaking about some kind of "optical illusion" in a diachronic perspective. In other words, frequency attenuation in the occurrence of core coordination units is registered on a macroscale over a longer time period for every hundred year's timespan, spanning from 1100 to the present; on a microscale continuous relative frequency fluctuations are recorded every ten years from 1800 until the present.

Unambiguous frequency curves alignment in macrodiachrony and the observed continuous frequency fluctuations in microdiachrony respectively demonstrate the result and process of discourse ability development. Discourse ability is understood as a set of certain rules to produce and interprete communicative behaviors (i.e. the communicative demand to encode and decode abstract relations among extralinguistic entities, in our case), solidified in texts of different genres in a particular social, cultural, and historic context. Discourse ability reflects speech behavior and thinking features of a given epoch, on the one hand, and various sociocultural spheres (literary, academic, official, etc. types of discourse), on the other [20: 204; 23: 145; 24: 73; 25].

The analysis above suggests that macrodiachronic attenuation of fluctuations in the occurrence of the units as frequent as coordination conjunctions reflects either the consolidating discourse production strategies or their eventual maturity. Meanwhile, the registered continuous fluctuations in microdiachrony, in particular the absence of the attenuation effect, testify to the functional continuity of these units. These curves reflect the process of discourse ability development, thus, justifying the statement by A. Meillet that coordination conjunctions «se renouvellent sans cesse» 'are continuously updated' [26: 9]. The constructed individual unit frequency curves demonstrate renewed demands to nominate additive, disjunctive, adversative, and causal relations between extralinguistic entities. Further research should give particular attention to the distribution of core coordinators by text genre at the stage of national language formation.

7 Conclusion

Our diachronic observations allow us to conclude that quantitative analysis of conjunctions as a high-frequency grammatical phenomenon should be aimed at striking the balance between macro- and microdiachrony, the time interval scale, and the frequency range. Consideration of these three parameters allows to establish individual usage trajectories for the selected core coordination units, including individual microdiachronic multidirectional trends in expressing core additive, adversative, disjunctive and causal relations between extralinguistic entities, as well as the overarching trend towards lower frequencies on macrodiachronic level.

In diachronic perspective individual usage development of every core coordination unit at comparable time intervals demonstrate that speakers' demands to nominate various segments of their experience undergo tranformations. Predominant *et* in the early period of language development and the higher frequency of *mais* in recent decades reflect cognitive and pragmatic significance of certain relations between extralinguistic entities across different periods of language history.

References

1. Luria, A. R. (1979). *Yazyk i soznaniye [Language and cognition]*. Izdatel'stvo Moskovskogo universiteta. (In Russian).
2. Kripke, S. (2001). *Naming and necessity*. Harvard University Press.
3. Putnam, H. (1975). *Mind, language and reality*. Cambridge University Press.
4. Lyons, J. (1995). *Linguistic semantics: An introduction*. Cambridge University Press.
5. Bybee, J. (2015). *Language change*. Cambridge University Press.
6. Godefroy, F. (2022). Dictionnaire de l'ancienne langue française et de tous les dialectes du IXe au XVe siècle. Retrieved 25 Jan. 2022, from http://micmap.org/dicfro/chercher/dictionnaire-godefroy/
7. Brunot, F. (1922). *La pensée et la Langue: Méthode, principes et plan d'une théorie nouvelle du langage appliquée au français*. Masson et cie.
8. Haspelmath, M. (2009). Coordination. Language typology and linguistic description (No. 2, pp. 3–50). Cambridge University Press.
9. Grevisse, M. (2018). *Le bon usage. Grammaire fraiņçaise*. Duculot.
10. Antoine, G. (1958–62). *La coordination en français* (Vol. 2). D'Artrey.

11. Abeillé, A. (2005). Les syntagmes conjoints et leurs fonctions syntaxiques. *Langages, 4*(160), 42–66.
12. Marchello-Nizia, C., Combettes, B., Prévost, S., & Scheer, T. (2020). *Grande Grammaire Historique du Français*. De Gruyter Mouton.
13. Dobrovol'skij, D. O. (2020). Corpus-based approach to phraseology research: New evidence from parallel corpora. Vestnik of Saint Petersburg University. *Language and Literature, 17*(3), 398–411. (In Russian).
14. McEnery, T., & Hardy, A. (2011). *Corpus linguistics*. Cambridge University Press.
15. Tognini-Bonelli, E. (2001). *Corpus linguistics at work*. John Benjamins.
16. Tummers, J., Kris, H., & Geeraerts, D. (2005). Usage-based approaches in Cognitive Linguistics: A technical state of the art. *Corpus Linguistics and Linguistic Theory, 1*(2), 225–261.
17. Frantext. (2022). National Corpus of French. Retrieved 20 Mar. 2022, from http://www.frantext.fr/
18. Baranov, A. N. (2001). *Vvedeniye v prikladnuyu lingvistiku [Introduction to Applied Linguistics]*. Editorial URSS. (In Russian).
19. Tarasevich, L. A. (2014). *Semantika i funktsionirovaniye predlogov s prostranstvennym znacheniyem (na materiale nemetskogo i russkogo yazykov)*. Minskiy gosudarstvennyy lingvisticheskiy universitet. (In Russian).
20. Baranov, A. N., et al. (2021). *Korpusnaya model' idiostilya Dostoyevskogo [Corpis model of Dostoevsky's idiostyle]*. Lexrus. (In Russian).
21. Auseichyk, Y. V. (2021). Semantika i funktsionirovaniye sochinitel'nogo soyuza *ou* v srednefrantsuzskom yazyke [Semantics and Functioning of the Coordinating Unit *ou* in Middle French]. Vestnik Moskovskogo gosudarstvennogo pedagogicheskogo universiteta. Filologiya. Teoriya yazyka. *Yazykovoye obrazovaniye, 3*(43), 40–53. (In Russian).
22. Auseichyk, Yu. V. (2021). Differentsiatsiya protivitel'nykh otnosheniy v semantike soyuza *mais* (diakhronicheskiy aspekt) [Differentiation of adversative relations in the semantics of *mais* (diachronic aspect)]. Vestnik Minskogo gosudarstvennogo lingvisticheskogo universiteta. *Filologiya, 1*(110), 77–87. (In Russian).
23. Fairclough, N. (1998). Political discourse in the media: An analytical framework. *Approaches to Media Discourse*, 142–162. Blackwell
24. Issers, O. S. (2012). Lyudi govoryat... Diskursivnyye praktiki nashego vremeni [People say… Discourse practices of our days]. Omsk: Izd-vo Omsk. gos. un-ta. (In Russian).
25. van Dijk T. A. (2008). *Introduction: Discourse and domination. Discourse and power* (pp. 1–26). Palgrave Macmillan.
26. Meillet, A. (1915). Le renouvellement des conjonctions. *École pratique des hautes études, Section des sciences historiques et philologiques*, 9–28.

The Use of *Futur Antérieur* in the Past in Old French: A Corpus-Based Study

Ekaterina Ivanova, Olga Khutoretskaya, and Maria Solovieva

Abstract The article discusses one of the secondary functions of the French future analytical form *futur antérieur*—completion in the past. The aim of the study is to examine how this secondary function appeared and developed in Old French. The study draws on the French epic texts of the eleventh-thirteenth centuries represented in the Ancien français subcorpus of the Frantext corpus. The corresponding corpus-based procedures resulted in creating a research subcorpus that includes 53 cases of the use of *futur antérieur* to express an action in the past. Comparative analysis of contexts with the meaning in question revealed signs of insufficient grammaticalization of *futur antérieur* in Old French. It allowed us to make a conclusion that the influence of the Latin possessive structure on the forming meaning of the analytical future created the prerequisites for reconsidering its temporal and aspectual meaning in the past. The conducted analysis confirmed the obligatory for Old French projection of the result of a past action expressed by *futur antérieur* to the future. Two types of reference to the future that appear on different levels of text organization have been defined.

Keywords Analytical future · Old French epos · Grammaticalization of the analytical form · Aspectual meaning · Result in the future

E. Ivanova · O. Khutoretskaya (✉) · M. Solovieva
St. Petersburg University, Universitetskaya emb. 7/9, 199034 St. Petersburg, Russia
e-mail: o.khutoretskaya@spbu.ru

E. Ivanova
e-mail: e.ivanova@spbu.ru

M. Solovieva
e-mail: m.solovieva@spbu.ru

P. Eismont et al. (eds.), *Literature, Language and Computing*,
https://doi.org/10.1007/978-981-99-3604-5_6

1 Introduction

Futur antérieur (hereinafter FA) or *futur composé* is a French analytical indicative tense-aspect form whose dominant functions refer to the future. Despite its relatively small use in the contemporary language this form is of ongoing interest to linguists. It can be accounted for by the fact that this form has a number of unique secondary functions besides the basic temporal meaning and researchers have not yet reached the agreement on its terminology and semantic characteristics. For example, a Polish researcher Ewa Ciszewska [1] offers the following list of the secondary functions of FA that were mentioned in works focusing on this tense: FA *épistémique, rétrospectif, exclamatif, impératif, de bilan, d'indignation, de protestation, historique, scénique, juridique, atténuatif (de cohésion).* Half of the names on this list reflect different nuances of the use of FA to express the past [2]. FA in the specified function is described in some works as an action that has finished in the present [3], i.e. due to its temporal and aspectual characteristics it can be compared to the French perfect *passé composé* (hereinafter PC). The interest in such use of FA has notably increased within the last decade because it tends to actively spread in the contemporary French media [4, 5]. That is why it is particularly relevant to clearly understand the peculiarities of this function as well as its work in the context and to correctly decode it.

The objective of the study is to examine the prerequisites for the formation of the FA ability to denote an action that takes place in the past in relation to the moment of speaking and to define the development of this meaning at the stage of formation of the semantic structure of the French tense in question.

2 Literature Review

The fact that the semantic structure of FA has a meaning of the past is mentioned in most French grammar books [6–9]. However, because of a rare use of this meaning the authors of these books tend to show little interest in it. They only observe the ability of FA to be transposed to the past and illustrate this ability by some linguistic examples taken mostly from classic French literature. In some grammar books the possibility of such transposition is not mentioned at all [10] or given as a remark [11].

The situation is different concerning specialized academic studies. Here disputes about the use of FA in contexts making it similar to *passé composé* has a long history. The main disputed questions address the following issues: specific peculiarities of FA having the meaning of the past; reasons making the speaker (or the author) use for this purpose a future grammatical form despite the variety of French past tenses; to what degree FA is equivalent in the specified function to the perfect tense PC. Two absolutely different points of view on all the above mentioned positions were expressed by French linguists Léon Clédat and Henri Yvon at the very beginning of the twentieth century. Based on linguistic examples from the Old French epos of the twelfth-thirteenth centuries and observations made by A. Meillet and J. Vendryes, Yvon [3] considers that the affective component of FA plays a crucial role in the meaning of the past. For him the primary thing at this stage of language development is not a distinct temporal reference of the action described by the grammatical form but the speaker's relation to it. Thus, the speaker's use of FA where PC is expected on the basis of the temporal order originates from the desire to convey emotional shades of meaning such as anger, hatred, disappointment, anxiety, indignation, satisfaction, pride, etc. These emotions are experienced by the speaker and brought about by the events expressed with the help of FA. The researcher thinks that temporal characteristics fade into insignificance and FA is equal to PC concerning temporal reference in such cases. L. Clédat holds an opposite opinion [12]. Developing Tobler's [13] conception, the researcher confirms that it is temporal characteristics of the FA form (its reference to the future and the aspectual meaning of completion) that form the basis for its transposition to the past. In Clédat's opinion, such transposition is possible provided that there is a special "projection" of the expressed action, event or condition into the future. What is meant here is an explicit or implicit "starting point" in the future which the speaker orients themselves to and from the perspective of which they consider the action as completed. Along with that, this action can precede, follow or happen at the moment of speaking. It is obvious that in this way functioning of FA in the past differs drastically from PC because the completion of the action expressed by PC refers to the speaker's present. In later works two approaches to the interpretation of the "retrospective" [6] function of FA were developed on the basis of the contemporary language. Modern researchers mainly approve of Clédat and his followers' point of view. They think that when FA is transposed to the past its "systematic" dominant temporal and aspectual characteristics do not change [14], it is "the interpretation of the starting point" that changes [15]. Some researchers consider that such change of the reference point is linked to the occurrence of some special meanings of FA, for example, evaluative modality [16], mediative meaning [15] or expressive coloring [17].

The mentioned above review of current interpretations proves that FA is a polysemic grammatical form. The ability to be transposed to the past with the aspectual meaning of completion already existed in the semantic structure of this form in Old French. That is why concerning the dominant functions of the studied grammatical form the following should be sought in Old French texts: reasons for the current use of FA to denote the past; the necessity for the specified meaning to refer to the future, answers to questions about its expressive component being primary or secondary and the possibility for FA to be replaced by PC. Those are the key tasks of this study.

3 Methodology

Data from the *Ancien français* subcorpus that includes eleventh-thirteenth century texts of different genres and the Frantext corpus that is a French language corpus are used in this study [18]. The research subcorpus consists of 59 documents with a total of 2 829 657 words dated 1 001–1299.

Firstly, a list of all possible spelling variants of Old French verb *aveir* in *futur simple* was made to get the research subcorpus of examples when FA is used to describe the past. This list includes 18 word forms given in dictionaries of Old French by Godefroy [19] and Van Daele [20]: *avrai, aurai, arai, auras, avras, aras, avra, aura, ara, avrons, aurons, avrium, aurez, avrez, avreiz, avront, auront, avrunt*. The uses of FA formed with the help of the auxiliary verb *estre* were excluded from this study because homonymous forms of this verb in future and imperfect (*er/ier*) do not allow to definitely interpret the analytical form as the future perfect tense (FA) or the past perfect tense (*plus-que-parfait*).

2012 word forms of the verb *aveir* in simple future found in the subcorpus of Old French were examined at the second stage of research. The examples containing homonymous forms, simple future and FA in its main meanings, i.e. used to describe the future, were deleted manually from the list. Thus, a research subcorpus that contains 53 cases of FA used for expressing an action in the past (the cases of FA used in its main meanings were examined apart from the research subcorpus in order to reveal "transitional" contexts contributing to the reconsideration of the main form).

The method of contextual analysis [21] was used while selecting examples for the research subcorpus and further work with the data obtained. The analysis of both the narrow and broad contexts of the use of FA in Old French allowed to monitor the conditions that created a favorable environment for the transposition of this form to the past in particular meanings.

4 Results

The contextual analysis of the cases with FA that were included in the research subcorpus as well as their further comparison with the examples of FA in its main meanings allowed to single out the following types of this form in Old French:

1. The primary function of FA in Old French is being an aspectual antagonist of *futur simple* and express completion of an action in the future—*le parfait du futur* (Clédat, 1927, p. 218)[1]:

(1) "Que vuels tu, frere? guarde n'i ait menti."

Et cil respont: "**Jel** vos **avrai** tost **dit**: …" (Le Couronnement de Louis, 1130, FR, p. 53)

In the given example there is a distant position of the auxiliary verb and the participle (*participle passé*) as part of the form of FA in the sentence. The limiting adverb *tost* ("at once, instantly, right now") shows that the action is limited in the future.

2. The meaning of anteriority over another action in the future that is formed on the basis of the aspectual meaning of completion.

[1] Hereinafter the examples from the Frantext corpus are used [18].

(2)—"Amis, bels frere", dist Gui li Alemans,
"Quant de Guillelme **avrai fine** le champ,
S'adonques vuelt icil suens nies Bertrans,
Ja por bataille n'en ira en avant." (Le Couronnement de Louis, 1130, FR, p. 77)

Guy the German says that at first he will finish the fight with Guillaume (*avrai fine*—FA), and after that he will be ready to fight with his nephew Bertrand. The future tense of the verb in the main clause (*ira—futur simple*) definitely localizes the whole situation in the future.

3. The meaning of a conditioned action in the future:

(3) "Si jo trai fors del feore ceste espee,
Ja vus **avrai** cele teste **colpee**!" (La chanson de Guillaume, 1150, FR, p. 105)

(4) Mestre, vus savez bien ke dit
Li sages homme en son repit:
De affaitement n'**avra** ja **pris**
Ki n'est fors de une cort appris. (HUE DE ROTELANDE, Ipomédon, 1180, FR, p. 73)

In examples (3) and (4) the meaning of completion in the future typical of FA overlays the meaning of conditionality of this action. As a result there is a distinct meaning of inevitability under certain conditions of the result in the future. Such meaning is contextually marked because it is expressed in the presence of a specific illocutionary meaning: a warning, a threat in example (3), cautionary or moral advice in example (4). The use of relevant vocabulary (teste colper—to behead; li sages homme—a wise man) contributes to the fact that the described situation gets a bright emotional coloring. In example (3) the condition that will have an inevitable result in the future is expressed by a conditional clause and apparently refers to the future. In example (4) the condition is not explicitly expressed but implied by the meaning of the subordinate clause ("the one who has not got manners will never get education"[2]). Such implicit condition is not directly related to the future, it is atemporal. In example (3) there is again a distant position of the participle in relation to the auxiliary verb as part of FA.

4. The meaning of an inevitable result in the future in relation to the protagonist (the subject of the denoted action). Such "transfer" to the past becomes possible because FA is used in the same function as the historical present that was often used to describe past events, which was typical of Old French texts:

(5) Car, quant il a un petitet juné,
Au celier vient si l'a tost desfremé,
Del pié le fiert si l'a tost enversé,
Vin vait querant tant qu'il en a trové,
De le vitaille tant qu'il en a assés.
S'on li desfent, mout tost l'**avra frapé**

[2] Hereinafter the translation is made by the authors of the article.

Ou par le pié a le paroi **jeté**. (Moniage Guillaume, 1150, FR, p. 10)

Example (5) illustrates a “transitional” context in which FA has its original meaning in atypical time. Such use of FA can be named as «*FA historique*». Example (5) is also illustrative from the point of view of the alternation of FA and PC. In this example there are actions repeated in the past in relation to the moment of speaking which is not the protagonist’s present: Guillaume comes up to the pantry and kicks the door open searching for food and wine. He does it till he finds enough food and drinks. If anyone disturbs him “he will immediately beat them or grab them by their leg and throw to the wall”. *Présent historique* and PC with the meaning of result in the protagonist’s present are used to denote repeated actions in the present. However, FA denotes inevitability of the result under certain circumstances in the perspective of the protagonist’s present-future, as in examples (3) and (4).

All the tense forms in example (5)—*présent*, FA and PC have a relative-temporal meaning. A distant position of the elements of FA and the presence of the limiting adverbs *mout* and *tost* that emphasize and strengthen the meaning of result should also be noted.

5. The meaning of result in the perspective of the future develops in the narrative of past actions in such a way that FA is reconsidered as a sign of completion of an action. The action objectively takes place in the past and its result is revealed and felt by the subject in the future. Two subgroups can be singled out among the examples found in the research subcorpus with such use of FA.

In the examples from the first group the reference of the result to the future has a more explicit and a more or less notable expression in the nearest surrounding context:

(6) Li cuens Guillelmes a sa gent apelee,
Tel chose dist qui a plusors agree:
“Or al harneis, franche gent onoree,
Si s’en ira chascuns en sa contree,
A sa moillier qu’**il avra esposee.**” (Le Couronnement de Louis, 1130, FR, p. 64)

(7) Renart, la male flame t’arde !
tantes foiz nos **avras folees**
et **chaciees** et **tribolees**…(Roman de Renart, 1180, FR)

In examples (6) an (7) the actions expressed by FA are located in the past and have been completed by the moment of speaking: warriors have already got married—example (6), Renard has already pursued and tortured his victims—example (7). However, the result of these actions is directly connected with the future. In example (6) the connection to the future is expressed by *futur simple* of the verb being a predicate in the main clause that precedes the verb in FA in the sentence. In example (7) the reference of the result to the future is implied by the subjunctive mood in the optative meaning (*t’arde*—«let it burn you»). In both examples the verb in FA expresses the meaning of summing up at the crucial point for the character or for the whole narration: the warriors get back to their wives in example (6), Renard has to meet King Lion's wrath for those sufferings that he inflicted on dame Pinte's relatives in example (7). One can assume that it is the meaning of result in the future that was provoked by some past process finished by the moment of speaking in combination with the semantic importance of the described moment that is reconsidered as summing up. It is obvious that this meaning is expressed in a bright emotionally coloured context against a background of the verbs of expressive semantics (*foler*—to fool, *chacier*—to pursue, *triboler*—to torture).

In the examples from the second subgroup FA denotes a process that takes place in the past and is completed at the moment of speaking without explicit references to the future in the nearest context:

(8) Dient qu'ils vienent d'Aufrique et d'outre mer
- Va donc, beau frere, lai les ceanz entrer;
Ge lor vorroie noveles demander,
Que fet mes sires, mout **avra demore**.»
Et cil lor cort la porte deffermer. (La Prise d'Orange, 1200, FR, p 0.61)

(9) Ahi! riche compaigne, comme or estes atainte!
Mors est cil qui des dones nos **avra faite** mainte. (PARIS (Alexandre de), Roman d'Alexandre, 1180, FR, p. 339)

Example (8) deals with a Saracenic king Thibaut who went to war with the French and did not come back for too long. While he was absent, Orange was conquered by Guillaume's army. The form of FA sums up the long absence of Thibaut who could have changed further events dramatically if he had been present. Thus, the action expressed by FA begins in the past in relation to the described situation and plays a crucial role for the further development of events.

In example (9) FA denotes an action that happens repeatedly in the past and stops at the moment of speaking due to objective reasons: the one who made gifts has died. A change of the situation influences significantly the characters and further events.

In the examples similar to (8) and (9) the meaning of summing up an action, event or a process taking place in the past comes to the fore in the semantic structure of FA. The absence of clear lexical or grammatical indications of the connection with the future is compensated by the importance of the obtained results for further development of the plot.

It can be said that in the examples of the first subgroup FA has a regular contextually marked meaning of a result in the future. It is often accompanied by the meaning of summing up and emotional colouring of the context. We defined collectively such meaning of FA as "result in the future". In the examples of the second subgroup the explicit reference of result to the future is absent. However, due to the importance of the described events for the development of the narration there is a regular meaning of summing up marked by the use of emotive vocabulary that adds emotional (mostly dramatic) colouring to the whole situation. We defined collectively such meaning as "summing up". Percentage distribution of two defined meanings of FA in the past in the research subcorpus is presented in Fig. 1.

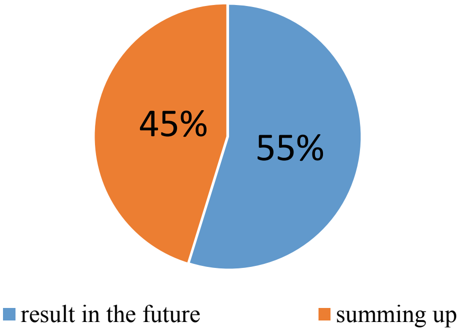


Fig. 1 Percentage distribution of the meanings of FA in the past (53 cases)

The chart shows that the meanings are distributed almost equally with a slightly higher number (10%) of the meaning of result in future. Thus, it cannot be said that any reference of a completed action to its result in the future (an explicit reference to the future in the nearest context or a projection of further development of the narration) is dominant while using FA in the past in Old French.

5 Discussion

FA with the auxiliary verb *avoir* etymologically goes back to the Latin “possessive” structure with a transitive verb of the *scriptum habēre habeo* type—«will have it written». In the process of becoming a grammatical analytical form the structure gradually passes from denoting the object’s feature to expressing an action completed by a certain moment in the future [22]. Many cases of a distant position of FA components in a sentence allow to speak about the incomplete grammaticalization of this form in the Old French period. If grammaticalization of FA in the considered period was not finished it may be possible to interpret this form as attributing a feature to the object and the object will have this feature in the future. What is more, the feature included in the predicative part which is expressed by the past participle must be already formed at the moment of being obtained by the object. Consequently, the action expressed by the participle starts in the past and finishes by the moment of obtaining in the future, which leads to the formation of the feature. The feature extends in time and the degree of its expression gradually increases to reach its maximum by a certain moment in the future.

Thus, due to its incomplete grammaticalization the FA form allows to connect under favourable conditions the expression of an action that takes place in the past with the result which the subject will use in the future. Such favourable conditions for this meaning were most likely formed at the point of the transposition of FA to the context with past events thanks to its use together with *présent historique* in the same time perspective. Later temporal and aspectual meanings of FA were gradually reconsidered in relation to the past.

The assumption that the transposition to the past took place when the grammaticalization of FA in Old French had not been finished is indirectly confirmed by the following fact. In all the examples found the meaning of the past is expressed by transitive verbs (*esposer, porter, asenbler, amer, vencre, pener, perdre, faire* etc.). Latin possessive structures with *habēre* comprised a past participle derived precisely from transitive verbs. The grammaticalization of the construction makes its components broader and more abstract and contributes to reconsidering its primary grammatical functions. After the initial semantic correlations were broken lexical units that had not been used in the construction before were involved in it. That is why the formation of FA with the meaning of the past expressed by intransitive verbs would have indicated a complete grammaticalization. The absence of these verbs suggests otherwise.

The analysis also showed that FA in the meaning of the past is used mainly when narrating crucial moments of the story or in the character's life. Such use appears to logically result from the ability of FA when being transposed to the past to sum up events (actions, conditions) that extend in time and to anticipate their consequences in the future. This use results in intensive emotions and dynamic description of the situation, which proves Yvon's observations about a strong expressive potential of the FA in the meaning of the past [3]. However, the expressive component is not dominant for the considered meaning of FA in Old French. The expressive and evaluative function can be considered as an accompanying meaning resulting from temporal characteristics of FA in the past. The needed effect is achieved thanks to lexical semantics of the verbs in FA and the use of intensifiers and limiting words.

Following Clédat, Taji and others [2, 12, 13], we cannot agree with Yvon's point of view [3] about the similarity of temporal and aspectual characteristics of PC and FA in the meaning of the past. Both PC and FA in the meaning of the past have an aspectual meaning of completion by the moment of speaking which turns to the meaning of result. However, PC is an aspectual antagonist to the present *(présent),* so the result of the action expressed by PC is felt in the speaker's present. However, the result of an action expressed by FA in the studied function reveals itself in the future. To illustrate this observation let us look at the example (10):

(10) De chrestïens devez estre servie.
Ne vos ait hume ki facet cuardie !
Mult larges teres de vus avrai cunquises, -
Que Carles tent, ki la barbe ad flurie,
E li empereres en est e ber e riches.» (La Chanson de Roland, 1125, FR, p. 176)

In example (10) dying Roland addresses his sword. The knight conquered vast lands for Charles the Great with this sword and the best knights of the kingdom will fight with this sword after the hero's death. The reference of the result to the future (the sword will go to other knights) is expressed by a combination of a modal verb *deveir* in the finite present form and the passive form of the infinitive. If knights fought with this sword at the moment of speaking, PC would be used to express all the preceding period when the sword had belonged to the hero. However, the hero is still alive and the change of the sword owner lies ahead. The moment when the process expressed by the analytical verb form is absolutely completed takes place in the future, which FA points at.

6 Conclusion

The analysis of the research subcorpus showed that in Old French the FA form, used in both the past and future often demonstrates a distant position of the components (the auxiliary verb in *futur simple* and the past participle *participe passé*). The observations made allow to assume that FA did not complete its grammaticalization as an analytical verb form in Old French. The incomplete grammaticalization appears to create preconditions for reconsidering FA and its use under certain circumstances to denote a process that begins in the past in relation to the moment of speaking and finishes (or has finished) at the moment of speaking having a result in the future. The transposition of FA with incomplete grammaticalization from the future to the past took place in several stages: the meaning of completion and an inevitable result in the perspective of the future through transitional relative-temporary context is reconsidered as a sign of precedence in relation to present-future and to the result in the same time period. Thus, FA denotes an action that was completed in the past with a result in the perspective of the present-future.

Thus, agreeing with Clédat [12] and his followers we came to the conclusion that in Old French FA in the past is not a direct equivalent of PC because it demonstrates an obligatory reference of the result to the future. However, the reference to the future is not achieved by transferring the "starting point" to the future. It is understood by the speaker due to the importance of the result of an action, process or a condition expressed by FA in the future. In addition to it, the conducted analysis showed that, in Old French epic texts the future reference of the result of an action expressed by FA can appear on different levels of text organization: the taxis level in poly-predicative constructions (the meaning of result in the future) and the text level as an occurrence of temporal order (the meaning of summing up).

The present study can be continued from the perspective of comparative analysis of FA functioning on different development stages of French and can be based on relevant subcorpuses of the Frantext corpus. The dynamics of the meanings of FA taken from Old French texts in different types of discourse of the contemporary language can be the subject for a further study.

References

1. Ciszewska, E. (2011). Comment peut-on déterminer les valeurs du futur antérieur? *Romanica Cracoviensia, 11*, 66–74.
2. Taji, K. (2003). À propos du futur antérieur. *L'information grammaticale, 97*, 37–40.

3. Yvon, H. (1922). Sur l'emploi du futur antérieur (futurum Exactum) au lieu du passé composé (passé indéfini). *Romania, 48*(191), 424–431.
4. Ciszewska-Jankowska, E. (2019). L'emploi du futur antérieur dans des textes de presse française. *Langue française, 201*, 115–130.
5. Skvortsova, A. D. (2013). Funktsii futur antérieur v sovremennom frantsuzskom yasyke [The functions of futur antérieur in modern French]. Vestnik Leningradskogo Gosudarstvennogo universiteta imeni A.S.Pushkina. *Seria Philologia, 2*(7), 181–189. (In Russian).
6. Wagner, R. L., & Pinchon, J. (1962). *Grammaire du français classique et moderne*. Librairie Hachette.
7. Imbs, P. (1968). *L'emploi des temps verbaux en français*. Klinoksieck.
8. Grevisse, M., & Goosse, A. (2007). *Le bon usage* (14^{e} éd.). De Boeck.
9. Wilmet, M. (1998). *Grammaire critique du français* (2^{e} éd.). Hachette Supérieur and Duculot.
10. Riegel, M., Pellat, J.-C., & Rioul, R. (2009). *Grammaire méthodique du français*. Puf.
11. Steinberg, N. M. (1972). *Grammaire française*. I. Prosveshchenie.
12. Clédat, L. (1927). Encore le futur antérieur. *Romania, 53*(209–210), 218–222.
13. Tobler, A. (1905). *Mélanges de grammaire française*. Trad. de M. Kuttner et L.Sudre. A. Picard et Fils.
14. Gosselin, L. (2019). Le futur antérieur d'un point de vue systémique. *Langue française, 201*, 31–46.
15. Apothéloz, D. (2019). À propos des emplois dits «passés» du futur antérieur. *Langue française, 201*, 61–77.
16. Ciszewska-Jankowska, E. (2014). *Le futur antérieur et ses emplois: analyse contextuelle*. Wydawnictwo Uniwersytetu Śląskiego.
17. Gak, V. G. (2000). *Teoreticheskaya grammatika frantsuzskogo yasyka [Theoretical Grammar of the French Language]*. Dobrosvet. (In Russian).
18. Frantext. Retrieved 19 Apr. 2022, from http://www.frantext.fr
19. Dictionnaire de l'ancienne langue française. Retrieved 12 Feb. 2022, from http://micmap.org/dicfro/search/dictionnaire-godefroy/
20. Petit dictionnaire de l'ancien français. Retrieved 12 Feb. 2022, from http://micmap.org/dicfro/search/vandaele-dictionary
21. Kolshanskiy, G. V. (1980). *Kontekstualnaya semantika [Contextual Semantics]*. Nauka. (In Russian).
22. Sabaneeva, M. K., & Shcherba, G. M. (1990). *Istoricheskaya grammatika frantsuzskogo yazyka [Historical Grammar of the French Language]*. Izdatelstvo Leningrandskogo Universiteta. (In Russian).

Nachhaltigkeit in Media Crisis Discourses

Irina Jesan, Elena Kovtunova, and Elena Sadovskaya

Abstract The present paper focuses on linguistic analysis of media crisis discourses presented in the modern German mass media. The approach that we took methodologically allowed us to determine the status of the lexemes *Nachhaltigkeit/nachhaltig* in the context of the crisis topic. The study specifically focused on the texts from the *Die Zeit* newspaper from 2008 to 2018 as they are presented in the DWDS corpus, as well as the texts from the online version of the newspaper found on its website from 2019 to January 2022. The study operates on a qualitative linguistic-discourse analytical research, a corpus-oriented method and a systematic lemma search method. The discourse analysis of the lexical units reveals its application in the period of the four relevant crises: the 2008 financial crisis and the following European debt crisis, the European migration crisis and the Corona crisis. Apart from the qualitative analysis of the frequency of use of the noun *Nachhaltigkeit* and the adjective *nachhaltig*, the paper illustrates other linguistic characteristics, such as its contextual semantic trends and collocations in the sociopolitical discourse in modern media.

Keywords Media discourse · Crisis discourse · Discourse studies · Corpus-based discourse analysis

1 Introduction

Discursivity as one of the characteristic aspects of communication has led to the increasing research interest to discourse as an object of study [Girnth 2012: 12]. According to Wengeler and Ziem, “socio-political and economic ‘crises’ […] undoubtedly belong to those ‘social facts’ whose half-life is closely related to the public discourse that systemizes them. As soon as they are no longer discussed in the

I. Jesan (✉) · E. Kovtunova · E. Sadovskaya
St. Petersburg University, Universitetskaya Emb. 7/9, 199034 St. Petersburg, Russia
e-mail: i.ezan@spbu.ru

P. Eismont et al. (eds.), *Literature, Language and Computing*,
https://doi.org/10.1007/978-981-99-3604-5_7

mass media, they disappear from (public) consciousness. And conversely, as soon as they become the focus of daily reporting, crises become ‘facts’ that become more confirmed with every day that they are discussed in the media” [12, 18: 52]. Recently, crisis discourses have been extensively studied by German researchers [14].

The study operates on a qualitative linguistic-discourse analytical research, a corpus-oriented method, contextual and semantic lexical analysis.

This study aims to investigate the lexemes *Nachhaltigkeit/nachhaltig* in German mass media crisis discourses. The qualitative discourse linguistic analysis implies the identification of *Nachhaltigkeit/nachhaltig* concordances in the context of the crisis topic in the *Die Zeit* newspaper in the DWDS newspaper corpus [7] and on the newspaper’s website [5], as well as its contextual semantic and distributive description (increased and decreased frequency of occurrence) and an attempt to determine the status of the lexeme *Nachhaltigkeit*. The corpus analysis was carried out by using analytical tools from the DWDS platform, most importantly collocations analysis. A lemma search method was also used in the current study.

It is important to mention general meanings of the lexeme *Nachhaltigkeit (sustainability)* found in the modern dictionaries. The first meaning we find in the German DUDEN dictionary is ‘längere Zeit anhaltende Wirkung’. The lexeme also has LSP-specific characteristics: in forestry—‘forstwirtschaftliches Prinzip, nach dem nicht mehr Holz gefällt werden darf, als jeweils nachwachsen kann’, in the field of ecology—‘Prinzip, nach dem nicht mehr verbraucht werden darf, als jeweils nachwachsen, sich regenerieren, künftig wieder bereitgestellt werden kann’ [6].

The Cambridge Dictionary gives the following definition: *sustainability*—Business English, Environment, Natural Resources: (1) the idea that goods and services should be produced in ways that do not use resources that cannot be replaced and that do not damage the environment, (2) the ability to continue at a particular level for a period of time [4].

The hypothesis is that especially in crisis discourses, *Nachhaltigkeit* demonstrates a strong thematic interest in the media public, particularly in the context of crisis management, because the demand for “long-lasting effect” of measures taken against crises is high. If the hypothesis is confirmed, the discourse keyword status of *Nachhaltigkeit/nachhaltig* in crisis discourses can be accepted. The term “*Schlüsselwort*” (*keyword*) is understood here in the discourse linguistic and political linguistic aspects as a word, which reveals a particular relevance in society for a certain period of time, promotes the aims and agendas of political institutions and is able to influence not only thoughts and feelings, but also people’s behaviour [16: 496].

2 Theoretical Principles

2.1 Mass Media Discourses. Media Crisis Discourse Study

We assume that provided topics and information that lead to the development of social discourses depend not on medial, but on mass medial distribution. Each mass medium has its specific characteristics, possibilities and limitations to establish and develop discourses: they differ according to the target group, the way information is conveyed, its functions, the degree of influence, etc. [8: 5]. “Mass media organize a public forum for discourses with a temporal and spatial scope” [13: 33].

The selection of topics, the type and intensity of its discussion cause thematic priorities among recipients and support a close connection between the individual and media-distributed relevance structures of interpretation and action, which correlates with the fact that mass media determine what, when and how is thematized, as well as which opinions are important [13: 33]. In this context media are concerned as mediating instances and actors at the same time, since they also produce offers of social reality interpretation and actively participate in the generation of discourse. Mass media underlie the processes of discursive construction: “through them, (potentially) transnational, trans-social discourse arenas are created” [2: 264].

Academic researchers describe leading media as the media that “present an explicit normative and relatively closed journalistic and editorial fundamental position” [1: 23]. Furthermore, they occupy one of the central positions in the public sphere and enable follow-up communication (Jarren and Donges 2006: 104). The leading media refer to ongoing opinion- and will-forming processes and influence social discourses [1: 23].

The article focuses on very specific discourses within the time period of 2008 to 2022 that address the financial, euro, migration and corona crises.

2.2 Nachhaltigkeit *in the Linguistic Research Area*

Though the study of *Nachhaltigkeit* belongs to one of the traditional fields of study in modern linguistics, it reveals a strong interest of academic community in the recent years. In 2017 an international workshop on “Nachhaltigkeit—Konzept, Kommunikation, Textsorten” was held at the Institute of German Philology in Greifswald, Germany [10]. As Gansel and Luttermann quite rightly emphasize, “the turn to linguistic reflection on the topic of sustainability communication and its structures is thus rooted in the fact that ‘Nachhaltigkeit’ is regarded as a guiding principle of the present and is a concern of society as a whole. That manifests itself in the language and communication” [10: 2].

The amount of texts dealing with *Nachhaltigkeit* has increased enormously over the last 20 years and the relative frequency of occurrence per million words has also risen. The frequency of occurrence of the lexeme raises the question of the meaning it has assumed over the past 30 years. With good reason it can be determined as a keyword due to its frequency of occurrence in terms of sustainability. The historical semantics describes keywords in terms of language use and develops the meaning in context usage. Collocations in sentence and text, aspects of linguistic action, thematic correlations of usage, the knowledge that the usage of keyword is assumed form types of an expression context usage [10: 2–3].

Gansel emphasises that Henn-Memmesheimer, Bahlo, Lubben and Qiu [11] quite reasonably criticize opinions that "the lexeme 'Nachhaltigkeit' has taken inflationary dynamics in the development and from this perspective can be characterized as a buzzword". Gansel goes on to explain that "frequency increases of word usage [...] do not per se point to a buzzword, especially since such lexemes are timely limited, which is precisely not the case with *Nachhaltigkeit*. Moreover, such factors as reduction, flattening of meaning or even emptying of meaning are not automatically associated with a frequency increase of word usage" [9: 46].

3 Analysis and Interpretation of Data

3.1 Study of the Corpus and the Key Factors of Its Actual Construction

This article examines a partial corpus from the *Die Zeit* newspaper. The basic criterion we defined for the analysis was the common communication context [3: 23], i.e. the explicit thematic focus on one of the three sociopolitical and economic crises mentioned since 2008. This study evaluates the three important crisis periods, as corresponded concordances determine three sub-corpora. The first sub-corpus contains *Die Zeit* publications on financial crisis of 2008 (publications between 2008 and 2009) and the following European debt crisis (publications between 2010 and 2011). These two crises can be merged into one single block related to the global financial problems around the world, including Germany and the European Union. The second sub-corpus contains *Die Zeit* publications related to the 2015–2016 European migration crisis. The *Die Zeit* corpus on the DWDS platform was used for the construction and analysis of the first and second sub-corpora. The third sub-corpus contains present publications from the online version of the *Die Zeit* newspaper and thematically covers the aspects of the Corona crisis within the time frame of 2020 to January 2022.

It is important to present the capabilities of the DWDS platform in more detail. "The aim of the project, which is based at the Berlin-Brandenburg Academy of Sciences and Humanities, is to create a "Digital Lexical System"—a comprehensive word information system accessible to users via the Internet, which provides information on German vocabulary in the past and present" [7]. The DWDS portal is much more than just a corpus, as it combines dictionaries, various corpora and statistical evaluations. Word frequency statistics, as well as course curve can be computer-based calculated. On the front page of the DWDS platform there is an extra section "text corpora", which is available for corpus-based linguistic studies. Using our corpus as an example, we can present the most important characteristics of the *Die Zeit* corpus (1946–2018). The analyzed corpus is classified as a newspaper corpus. There are indications of sentences and tokens. The texts can be searched for free or after registration. The entire *Die Zeit* corpus is freely accessible and contains 1,212,177 documents consisting of 30,486,227 sentences and 563,279,363 tokens.

The frequency of *nachhaltig* and *Nachhaltigkeit* was computer calculated in the DWDS newspaper corpus. The frequency of the words in the *Die Zeit* newspaper was calculated as: in 2008–2009 *nachhaltig* (923 references), *Nachhaltigkeit* (222 ref.); in 2010–2011 *nachhaltig* (1873 ref.), *Nachhaltigkeit* (394 ref.); in 2015–2016 *nachhaltig* (1685 ref.), *Nachhaltigkeit* (309 ref.). Over the whole time period we see the dominance of the adjective above the noun in terms of frequency of use.

3.2 Media Crisis Discourse 2008–2012 (Economic Crisis)

On 15 September 2008, the investment bank Lehman Brothers, which had been heavily involved in the real estate loan business, filed for bankruptcy—with far-reaching consequences. The financial crisis reached its peak. The German economy shrank by about five percent in 2009. Since 1946, there had only been five years with negative economic growth in Germany. Development in various areas of the financial markets and the economy contributed to the devastating extent of the 2008 financial crisis.

The European debt crisis (also known as the eurozone crisis or euro-crisis in German) is a multi-layered crisis of the European Monetary Union that began in 2010. This crisis includes a sovereign debt crisis, a banking crisis and an economic crisis. The term does not refer to the external value of the euro, which remained relatively stable.

Based on the fact that texts are connected with a common topic within the framework of a discourse, which is usually controversial and therefore has to be negotiated, the discussed facts are evaluated positively or negatively and justified argumentatively [Girnth 2012: 12–14]. In the following example, we can see how people can deal with the term *Nachhaltigkeit* in a relatively critical and skeptical way:

> Wem Zu Den Globalen Krisen Gar Nichts Mehr Einfalle, Der Rede Von Nachhaltigkeit. (Die Zeit, 12.11.2009, Nr. 47)

On the other hand, most contexts show the positive effects of a sustainable approach in the economic crisis situations:

> Unternehmen denken in der Krise mehr über Nachhaltigkeit und gesellschaftliche Verantwortung nach, auch weil sie wieder das Vertrauen ihrer Stakeholder gewinnen wollen, sagt Faruk optimistisch. (Die Zeit, 09.03.2009, Nr. 11)

The metaphor based on the concept *the way* enjoys great popularity in crisis contexts:

> Um eine Aufwärtsbewegung in Gang zu setzen, müsste die Politik einen nachhaltigen Ausweg aus der Krise aufzeigen—doch der ist nicht in Sicht. (Die Zeit, 14.07.2008, Nr. 29)

An efficient economy also includes sustainability, as the following example demonstrates:

> Statt durch ihre vielen Milliarden eine neue, nachhaltigere Wirtschaftsweise zu fördern, gibt sie sie für übliche Infrastrukturprojekte aus und hilft zusätzlich durch die beschlossene Abwrackprämie einer Autoindustrie, die ihre Krise zumindest teil-weise selbst verschuldet hat. (Die Zeit, 13.01.2009, Nr. 3)

The following collocations with *nachhaltig* were found in the publications related to the crisis: *ein nachhaltiges Wachstum; eine nachhaltige Entwicklung, Steuerung, Verbesserung, Wirtschaftsweise, ein nachhaltiger Ausweg, Pfad.*

The adjective *nachhaltig* is used in the contexts with the following verbs: *nachhaltig engagieren, modernisieren, sichern, wahrnehmen, verschärfen* (the last verb being used with a negative meaning).

The article published on 25.03.2011 provided the information on the frequency of occurrence of the two lexemes *Nachhaltigkeit* and *sustainability* in Google search results:

> 2,75 Millionen Treffer Bei Google Für "Nachhaltigkeit", Das Englische „Sustainability " Schafft Sogar Fast 30 Millionen. (Die Zeit, 25.03.2010, Nr. 13)

One newspaper article on the topic of the sovereign debt crisis discusses the ways to solve the acute problems, taking into account the important role of sustainability:

> Aber wenn es sich bei der aktuellen Krise im Kern um eine Staatsschuldenkrise handelt, dann wird sie nur gelöst werden können, wenn die Politik umsichtig und nachhaltig agiert (Die Zeit, 27.10.2011, Nr. 44)

Within analyzed time frame, only the adjective *nachhaltig* occurs in the contexts that describe the economic crisis. The collocators of *nachhaltig* are verbs, such as *agieren, umsteuern, vertreten, bekämpfen.* The reflexive verb *sich erholen* occurs in the following context focused on the global economy:

> Noch habe die Krise nicht auf aufstrebende Nationen und Entwicklungsländer übergegriffen—und alles deute weiter darauf hin, dass die Weltwirtschaft sich nachhaltig von ihrer schwersten Rezession seit sechs Jahrzehnten erhole. (Die Zeit, 08.07.2010)

The DWDS portal offers to use the tool DiaCollo for statistical analysis (collocation analysis in diachronic perspective). The idea is that you can automatically calculate the collocations for different time periods and see how they change over time. The *Die Zeit* corpus was chosen and the lemma *nachhaltig** was typed into the query form. The generated HTML collocation table represents the collocation change for the time frame of 2008–2012. In the table we can see collocators that occur significantly more often together with the search term *sustainable*. The most significant collocators are as follows: in 2008 *Entwicklung, Lebensstil, Wirkung, Lösung, Gewinn (development, lifestyle, impact, solution, profit);* in 2009 *Wachstum, Charta, Entwicklung, Erholung, Aufschwung, Geldanlage, Wirtschaft, verändern* (*growth, charter,* *development, recovery, upswing, investment, economy, change);* in 2010 *Wachstum, Entwicklung, ausgewogen, ökologisch, Geldanlage, Tourismus, Mobilität* (*growth, development, balanced, ecological, investment, tourism, mobility);* in 2011 *Wachstum, Landwirtschaft, Energieversorgung, Mobilität, Entwicklung, Liberalismus, schaden (growth, agriculture, energy supply, mobility, development, liberalism, harm);* in 2012 *Wachstum, Entwicklung, Mobilität, Wirkungsmöglichkeit, stören, beeinträchtigen, profitabel (growth development mobility, opportunity, disrupt, affect, profitable).*

We could suppose that a content-related overlapping can occur by comparing the results of discourse qualitative and automatic corpus-based linguistic approaches in the process of the collocations analysis.

Among the above-mentioned words, on the one hand, there are some words with more general semantics that can be used in different crisis contexts: *Entwicklung, Wirkung, Lösung*. On the other hand, since 2009 lexemes with a strong economic semantics that are typical for financial and economy contexts have become more frequent, for example, *Aufschwung, Geldanlage, Wirtschaft, Wachstum, profitabel.*

3.3 Media Crisis Discourse 2015–2016 (Refugee Crisis)

The refugee crisis was one of the most discussed topics in the sociopolitical discourse. Its representation in the media is related to reality because the cognitive perception of reality depends largely on the versions and interpretations given and distributed in the mass media. Numerous research publications have been dedicated to this political and social event [cf. 15].

Although the frequency of occurrence of the noun *Nachhaltigkeit* (309 references) and the adjective *nachhaltig* (1685 ref.) is traditionally high, especially within this time frame, compared to the time periods presented earlier, contexts related to crises with sustainability/sustainable can only be found sporadically.

In 2015 and 2016, a sustainable solution to the refugee crisis is often discussed and different actors are placed in the spotlight:

> Verpflichtende Quoten und die Verteilung von Flüchtlingen gegen ihren Willen sind keine nachhaltige Lösung der aktuellen Krise", sagte der tschechische Ministerpräsident Bohuslav Sobotka. (Die Zeit, 27.05.2015, Nr. 21)
>
> Allerdings darf es nicht den Blick dafür verstellen, dass Deutschland allein keine nachhaltige Lösung der Krise erreichen wird. (Die Zeit, 06.11.2015)

Next year there were already publications which focused on the lasting threat and discussed the possible ways to reduce the number of refugees. This is where *the water* concept-based metaphor comes into play *(Schockwellen, ein beispielloser Strom von Flüchtlingen):*

> Die Schockwellen, die von einer weiteren Eskalation der dortigen Krisen ausge-hen, stellen für den deutschen Außenhandel eine nachhaltige Bedrohung dar. (Die Zeit, 08.01.2016).

The comprehensive policy measures are discussed at different political levels:

Begrüßt wird das "sehr starke Engagement", die Krise durch ein umfassendes Pa-ket an internationalen, europäischen, regionalen und nationalen Maßnahmen zu lösen: "Ziel ist es, die Zukunftsaussichten und Lebensbedingungen der Menschen in Krisenregionen zu verbessern und den beispiellosen Strom von Flüchtlingen signifikant und nachhaltig zu reduzieren" (Die Zeit, 05.04.2016).

By using the tool DiaCollo besides the tabular visualization format, there are other graphic visualization options: Cloud, Bubble, etc. Figs. 1 and 2 illustrate the 10 frequent collocators of *nachhaltig** as Bubbles at the beginning of 2015 and the end of 2016 generated by DiaCollo (DWDS), which once again shows some dynamics in the lexical content of sociopolitical debates in this time period.

Of particular interest is the appearance of the lexeme *Flüchtlingszahl* at the end of 2016, as well as the acoustic abbreviation *SDG*. "At a UN summit in September 2015, the "2030 Agenda for Sustainable Development" was adopted. It should contribute to raising the standard of economic, environmental and social sustainability. All 193 member states of the United Nations committed themselves to implementation at regional, national and international level. This agenda identified 17 Sustainable Development Goals—the 17 Sustainable Development Goals (SDGs)" [17].

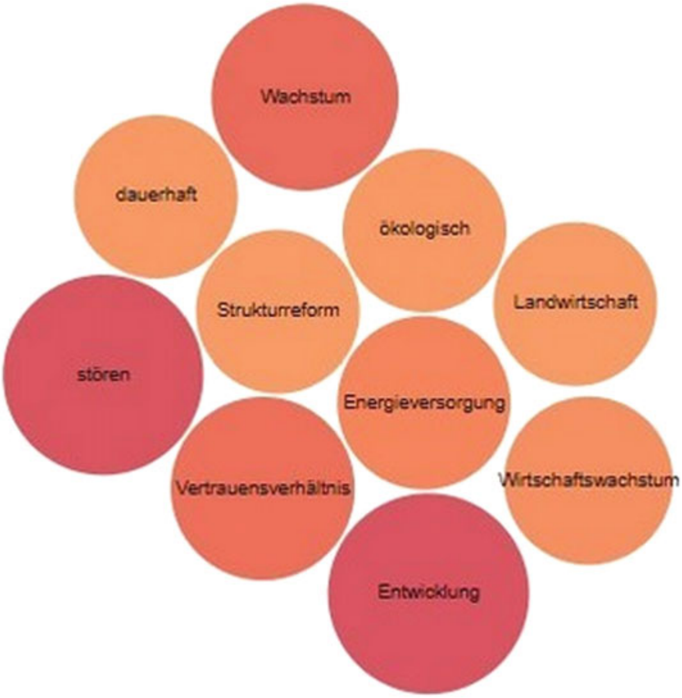


Fig. 1 DiaCollo (collocations) *nachhaltig** 2015

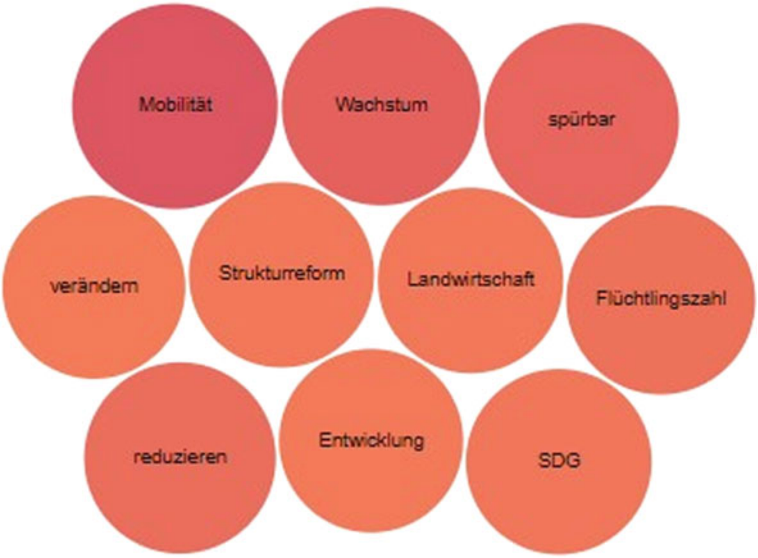


Fig. 2 DiaCollo (collocations) *nachhaltig** 2016

3.4 Media Crisis Discourse 2020–2021 (COVID-19 Pandemic)

The COVID-19 pandemic (coronavirus pandemic) was described in the German mass media also as a crisis, which currently revealed the health and medical dimensions along with its economic, financial, social and political criteria.

The metaphor of struggle, which is also typical for other types of crises, is expressed here to determine a resistance of the new and the old, whereby the analyzed terms *Nachhaltigkeit/nachhaltig* refer to the future and contextualize its positive connotation:

> Auch diese Krise ist, unter anderem, ein Ringen des Alten mit dem Neuen, das sich weniger im Wesen der Krise zeigt und mehr in den Reaktionen und in der Perspektive auf das, was danach kommt. Das Hoffnungspotenzial liegt in den Lösungen und Antworten, die den Status quo verändern. Eine andere, bessere Zukunft scheint vielen auf einmal möglich, mit mehr Solidarität, Nachhaltigkeit, einer grundsätzlich neu gedachten Gesellschaft. Aber die Hüter des Gestern sind hartnäckig (Die Zeit 10. April 2020).

As stated in the abstract above, besides the corona pandemic there are several crises that people are confronted with. The lexemes like *Fronten* point to the *war* concept-based metaphor (crisis is here and now, we have to fight it), but *Nachhaltigkeit* is thematised in the direction "what comes after the crisis", accompanied by the lexemes and phrasemes with positive evaluation *erfolgreich, Chancen erhalten/vergrößern, wichtig, zukunftsorientiert:*

> Die skizzierten Krisen gegeneinander auszuspielen, kann und darf unser Handeln nicht bestimmen. Die Herausforderung besteht darin, an den drei Fronten zugleich erfolgreich zu sein, um gut durch die Zwanzigerjahre zu kommen und Chancen für Veränderungen zur Nachhaltigkeit zu erhalten oder gar zu vergrößern. […] Dieser Reflex erschwert es jedoch oft, wichtige, zukunftsorientierte Innovationen—beispielsweise in Richtung Nachhaltigkeit—anzugehen. Krisen können cognitive lock-ins, Verharren in Vergangenheitsstrukturen, auslösen. Deshalb braucht es starke Stimmen, die zeigen, wie Zukunftsinvestitionen so gestaltet werden können, dass sie auch die Ängste und Verunsicherungen im Hier und Jetzt adressieren. (Die Zeit 2. April 2020).

The next media text contrasts different images of the future after the Corona crisis where *die Digitalisierung* is described with more positive semantics than *die Nachhaltigkeit:*

> Etwas Hoffnung gibt es: In einer Umfrage der Stiftung Familienunternehmen unter rund 500 Vertreterinnen und Vertretern von Deutschlands Familienunternehmen zwischen 16 und 39 Jahren sah jeder Dritte in der Digitalisierung Chancen für die eigene Firma—deutlich mehr als in einem Strategiewechsel oder in einem Ausbau der Nachhaltigkeit. (Die Zeit 17. Oktober 2020).

In the media pandemic discourse, *Nachhaltigkeit* is also given the seme "what people wish" and also co-occurs with the lexemes *Zusammenhalt, Klimawandel, Frieden und Solidarität:*

> In diesem Jahrhundert müssen Menschen das bewältigen, was in den vergangenen Jahrzehnten auf der Strecke geblieben ist: Klimawandel, Frieden und Solidarität. Auch in den Umfragen sehen wir, dass sich die Befragten mehr Nachhaltigkeit und Zusammenhalt wünschen. (Zeit-Magazin, 2. Februar 2021)

However, in the following publication, which describes a family story in the times of the Corona, *nachhaltig* and *digitalisieren* stand semantically positively side by side and support each other:

> Michael Otto hat Deutschlands größten Versandhändler nachhaltig gemacht, als noch alle darüber lachten. Er digitalisierte seinen Familienkonzern und schickte den jungen Jeff Bezos weg. Interview mit einem sturen Erfolgsmenschen. (Die Zeit 14. Mai 2021)

The following context illustrates the public debate on financial and political strategies of crisis management. *Die Nachhaltigkeit* is again surrounded by lexemes with positive connotations *(Verantwortung, Allgemeinwohl),* but this time it fights the negative contextual counter-terms from the field of economics and finance: *Abkehr vom Wirtschaftswachstum, die Schuldenlast tilgen:*

> Sollte Finanzminister Gernot Blümel Schwabs gleichnamiges Buch gelesen haben, konnte er aufatmend feststellen, dass darin zwar viel von Verantwortung, Nachhal-tigkeit und Allgemeinwohl die Rede ist, aber keine allgemeine Abkehr vom Wirtschaftswachstum gefordert wird. Denn Letzteres brauchen Blümel und seine Amtskollegen, um die Schuldenlast irgendwann wieder zu tilgen. Ein Konjunkturaufschwung bleibt auch nach einem großen Systemwechsel das wirksamste Mittel gegen die Folgen einer Krise. (Die Zeit, 16. Januar 2021)

Die Zeit featured a guest article by Niklas Potrafke "*Wann sind Staatsschulden richtig? Und wann sollte damit Schluss sein?*" in which possible financial solutions of the coronavirus crisis were discussed. The lexeme *nachhaltig* used in the lead updates its original climate-bound semantics and actualizes its new economy-specific aspects:

> Warum sich die junge Generation neben dem klima auch um nachhaltige Finanzen sorgen sollte. (Die Zeit, 17. Oktober 2020)

In our research we have observed that the lexemes *Nachhaltigkeit/nachhaltig* mostly receive a positive connotation when used in media texts related to the pandemic crisis. Furthermore, we have identified a separate group of examples where lexemes *Nachhaltigkeit/nachhaltig* were used in co-occurrence with negative related vocabulary, so that they updated occasionally the negative semantics:

> Der Corona-Shutdown ist dabei, ganze Volkswirtschaften nachhaltig zu ruinieren. Entlassungen, Kurzarbeit, Insolvenzen. Immer mehr Menschen wissen nicht, wie sie noch die nächste Miete bezahlen sollen. (Die Zeit 20. Juni 2020)

> Kürzlich hat ein großer Discounter das Kilo Schweinefleisch um einen Euro teurer verkauft, damit die Bauern höhere Einnahmen haben und mehr für die Tiere tun können. Dieser eine Euro war den Kunden zu viel, ihnen war Nachhaltigkeit das nicht wert. Die Idee ist gescheitert. (Die Zeit 14. Mai 2021)

> Im Prinzip gibt es zwei Arten, mit sich ankündigenden Krisen umzugehen: entweder zu warten, bis die Not etwas diktiert—oder Verantwortung zu übernehmen, auch wenn dann eine Minderheit zorniger Wähler noch zorniger wird. Das ist bei Corona nicht anders als beim Klima oder auch der Biodiversität. Variante eins haben die Deutschen nun schon des Öfteren ausprobiert. Sie hat nicht funktioniert, nicht bei der Nachhaltigkeit und schon gar nicht bei der Pandemie, bei der frühes Handeln von Anfang an das oberste Erfolgsrezept gewesen ist. (Die Zeit, 26. Juli 2021).

4 Conclusion

As can be seen from the analysis results, the noun *Nachhaltigkeit* and the adjective *nachhaltig* are relevant in the crisis discourses and have a high frequency of occurrence. In comparison with the noun, the use of adjective plays a dominant role. Both lexemes contribute to the construction of social crisis reality by primarily pointing to positive ways of solving it and by bringing a positive stimulus into the social debates. The discourse keyword character of these lexemes makes them recognizable and allows media texts to be distributed to a wider audience.

A certain combination of the linguistic methods applied in this research (discourse linguistic lexical-oriented and corpus-based analysis with contextual semantic analysis of lexis) allowed us to gain a more detailed understanding of current lexical, as well as general linguistic processes in German.

The integration of these combined methods provided a clear visualization of scientific concepts and observations. Moreover, it can be applied to conduct a more detailed analysis of different lexical layers in a discourse considering extra-linguistic factors.

References

1. Blum, R., Bonfadelli, H., Imhof, K., & Jarren, O. (Eds.). (2011). *Krise der Leuchttürme der öffentlichen Kommunikation.* Verlag für Sozialwissenschaften.
2. Brand, A. (2012). *Medien—Diskurs—Weltpolitik: Wie Massenmedien internationale Politik beeinflussen.* Trascript Verlag.
3. Busse, D., Teubert, W. (1994). Ist Diskurs ein sprachwissenschaftliches Objekt? Zur Methodenfrage der historischen Semantik. In: D. Busse, F. Hermanns, W. Teubert (Eds.), Wolfgang (Hg.): Begriffsgeschichte und Diskursgeschichte. Methodenfragen und Forschungsergebnisse der historischen Semantik. S. (pp. 10–28). Westdeutscher Verlag.
4. Cambridge Dictionary, https://dictionary.cambridge.org/dictionary/english/sustainability, last accessed 2022/05/2.
5. Die Zeit, https://www.zeit.de, last accessed 2022/05/2.
6. DUDEN, https://www.duden.de/rechtschreibung/Nachhaltigkeit, last accessed 2022/05/2.
7. DWDS—Digitales Wörterbuch der deutschen Sprache, https://www.dwds.de/, last accessed 2022/09/12.
8. Fraas, C., Klemm, M. (2005). Diskurse—Medien—Mediendiskurse. Begriffsklärungen und Ausgangsfragen. In: C. Fraas, M. Klemm (Hg.) (Eds.), Mediendiskurse. Bestandaufnahme und Perspektiven. S. 1–8. Peter Lang, Frankfurt/M (2005).
9. Gansel Ch. (2021). Nachhaltigkeit und der gesellschaftliche Resonanzraum. Gansel, Ch., Luttermann, K.: Projekt Angewandte Linguistik: PAL 1 In: Nachhaltigkeit—Konzept, Kommunikation, Textsorten (pp. 45–81). LIT Verlag.
10. Gansel Ch., Luttermann K.: Einführung. Gansel, Ch.; Luttermann, K. (2021). Projekt Angewandte Linguistik: PAL 1 In: Nachhaltigkeit—Konzept, Kommunikation, Textsorten (pp. 1–9). LIT Verlag.
11. Henn-Memmesheimer, B., Bahlo, Ch., Eggers, E., Mkhitaryan, S. (2012). Zur Dynamik eines Sprachbildes: Nachhaltig. In: R. Hansen-Kokoruš, B. Henn-Memmesheimer, G. Seybert (Hg.) (Eds.), Sprachbilder und kulturelle Kontexte. Eine deutsch-russische Fachtagung. Mannheimer Studien zur Literatur- und Kulturwissenschaft (Bd. 50, S. 159–187). St. Ingbert.
12. Kämper, H. (2017). Personen als Akteure. In: Roth, K.S., Wengeler, M., Ziem, A. (Hg.) Handbuch Sprache in Politik und Gesellschaft. De Gruyter.
13. Konerding, K.-P. (2005). Diskurse, Themen und soziale Topik. In: Fraas, C., Klemm, M. (Hg.) (Eds.), Mediendiskurse. Bestandaufnahme und Perspektiven. S. 9–38. Peter Lang.
14. Kuck, K. (2018). Krisenszenarien. Metaphern in wirtschafts- und sozialpolitischen Diskursen. De Gruyter.
15. Luft, S. (2016). *Die Flüchtlingskrise: Ursachen, Konflikte.* Folgen. C.H. Beck Wissen.
16. Niehr, Th. (2007). "Schlagwort". In: Historisches Wörterbuch der Rhetorik. Ueding G. (Hg.) (Eds.) in 10 Bdn. Bd. 8: Rhet — St. Tübingen: Niemeyer. S. (496–502).
17. Sustainable Development Goals—ESG Cockpit (esg-cockpit.com)/last accessed 2022/09/12.
18. Wengeler, M., Ziem A. (2014) Wie über Krisen geredet wird. Einige Ergebnisse eines diskursgeschichtlichen Forschungsprojekts Martin. *Zeitschrift für Literaturwissenschaft und Linguistik* 44, 52–74.

Using Corpora for Verifying Language Choices in Translation

Adelya Kh. Abdulmanova, Ekaterina K. Vyunova, and Irina A. Lekomtseva

Abstract The article focuses on how we can use corpora for translation purposes, i.e. for verifying cross-linguistic correspondences in translation in terms of the communicative equivalence. Correspondences in translation can be systematic cross-linguistic correspondences, yet they may be not natural and common for the target language as, according to the results of the corpus-based analysis, the co-occurrence relations between the words are not natural or common for the target language. Such collocation errors are largely due to approaching translation as transformation of language forms from the source language to the target language, rather than speech production in the target language, which lies within the frame of the activity theory. In translation, the focus should shift from language system to language use to achieve naturalness in translation. To this end, as corpora reflect language use, the article outlines some possibilities corpora have to offer to translation studies and contrastive analysis for translation purposes.

Keywords Cross-linguistic correspondences · Corpus-based methods · Communicative equivalence in translation

1 Introduction

With the advent of corpus-based and corpus-driven technologies into linguistics and translation studies, we can gain a deeper insight into the fundamental categories of translation studies. Among the fundamental categories of translation studies is the category of equivalence [9]. It can be defined as 'a specific relationship between texts that makes it possible to consider one text as a translation of another' [2, p. 26]. Even a brief overview of what translation equivalence is shows that what underpins the essence of and how to assess the quality of the specific relationship between the source text and the target text are dramatically different across various approaches [21]. Our approach to defining the category of equivalence as a specific relationship between

A. Kh. Abdulmanova · E. K. Vyunova · I. A. Lekomtseva (✉)
St. Petersburg University, Universitetskaya Emb. 7/9, 199034 St. Petersburg, Russia
e-mail: i.lekomtseva@spbu.ru

P. Eismont et al. (eds.), *Literature, Language and Computing*,
https://doi.org/10.1007/978-981-99-3604-5_8

the source text and the target text presupposes a linguistic approach to translation studies. Its essence lies in revealing cross-linguistic correspondences at different language levels. Cross-linguistic correspondences traditionally underlie the concept of translation equivalence. In this regard, since preserving the meaning in the target text is essential, the criterion for translation accuracy, or sameness, is the identity, or sameness, of the information conveyed by the linguistic forms in the source and target texts [22]. Only the translation that conveys this information by equivalent cross-linguistic correspondences can be considered as accurate or equivalent translation. In this regard, a curious and simple question may arise: what are the equivalent cross-linguistic correspondences in translation? The answer is seemingly simple, too. Equivalent correspondences in translation are the linguistic forms that are natural, common forms in the target language as regarded within the paradigm of natural equivalence by A. Pym that is heavily derived form J.P. Vinay and J. Darbelnet [21, p. 34]. In Russian translation studies, these natural forms are regarded as standard/regular/typical forms of expressing the original meaning in the target language [23]. Yet, on the level of practice, it is far from being easy to identify and verify such cross-linguistic correspondences as translation choices, especially in terms of the communicative equivalence, as the focus should be on language use, rather than on language system.

Interestingly, the problem of conveying the meaning in the target language is traditionally considered in the light of cross-linguistic correspondences and translation transformations, techniques [4]. This approach presupposes analysing the source text and its translation/translations to identify systematic cross-linguistic correspondences and translation techniques. Such an approach views translation within the frame of the substitutive-transformational paradigm [10]. This means that translation is mostly about optimising our search of translation transformations and, to a greater or lesser extent, is a process of ‘assembling’ the target text from the ‘forms’ of the source text by substitutions, rearrangements, additions (i.e. known in the Russian translation theory as translation techniques, or transformations) [10].

The substitutive-transformational approach will be opposed to translation as an activity, i.e. activity type of translation ontology [10]. This approach to translation is within the frame of the activity theory developed by Alexei Leontiev [13], Lev Vygotsky [31], etc. Translation is not manipulation/transformation of various linguistic forms across languages, but a speech production in the target language according to the programme of the source text [10]. If we approach translation as a speech production, our main task is to identify how linguistic forms are used in speech in the source language and the target language in contrast since translating activates translator’s competence both in the source language and the target language. The focus shifts from language system to linguistic use in contrastive perspective. When translating, it is necessary to find those cross-linguistic correspondences that are natural, common, idiomatic, regular means of expressing the given meaning in the source language and in the target language [23].

Our analysis of translations shows that on the level of practice there is a tendency towards viewing translation within the frame of the substitutive-transformational paradigm [11]. In other words, translators tend to use regular systematic cross-linguistic correspondences, i.e. those that can be found in the bilingual dictionaries. Verifying such language choices is rather simple. What we should do is to consult bilingual dictionaries or explanatory dictionaries to conduct a comparative seme, or componential, analysis. These systematic cross-linguistic correspondences are correct, adequate, or equivalent from the point of view of systematic relations between languages in contrastive perspective. Yet translation is not about the abstract language system/subsystems, but language use and speech production. Therefore, identifying the patterns of how language forms are used in speech and their natural patterns of co-occurrence in contrastive aspect is essential. In other words, what we should identify is correspondences between the usage of language units/forms in speech [25] in order to avoid features of the source language in the target text. Verifying how possible cross-linguistic correspondences are used in speech in the source language and in the target language is a hard nut to crack. To this end, we should verify the correctness of usage of language units/forms in speech. The focus shifts from bilingual or explanatory dictionaries to texts, i.e. from language system to language use. More often than not, verifying language choices in translation in such perspective is subjective. You can hear such phrases as ‘it sounds wrong’, ‘not everything we find in texts is always “natural” or “common”’ etc. [6]. However, such statements are far from being objective and depend on individual perception. To identify and objectively analyse the patterns of using language forms in speech in the comparative aspect has become possible largely after incorporating corpus-based methods into contrastive analysis and translation studies as corpora reflect language use [3, 5, 14].

The purpose of this article is, then, to analyse and verify language choices in translation in terms of their communicative equivalence. Such studies seem to be much up-to-date largely due to the fact that the results obtained can be used both in the theory of translation, i.e. in gaining a deeper insight into such basic categories of translation studies as translation correspondences, translation norms; and in translation quality assessment.

2 Methodology

The purpose of the article is to conduct a contrastive analysis in order to verify language choices in translation in terms of whether they are communicatively equivalent correspondences, or natural equivalents, using corpus data. To achieve this aim, it is necessary to solve two main tasks: first, to determine the relationship between translation correspondences at the level of the language system and, second, at the level of language use.

The following research methods were used: the method of seme, or componential, analysis; the method of corpus analysis of lexical units; and the comparative method. The analysis was carried out according to the following scheme. First, a comparative analysis of translation correspondences from the point of view of systematic relations was carried out using the method of seme, or componential, analysis of lexical units based on the data from explanatory dictionaries and/or translation, bilingual dictionaries [29]. Second, a contrastive analysis of translation correspondences was carried out in terms of analysing the patterns of their usage in speech, especially analysis of co-occurrence relations using the corpus methods. In our study, we use the National Corpus of the Russian Language (NCRL) [15]. We use the NCRL in two ways. First, to search for a phrase in a specific grammatical form, we use the field 'Search for exact forms'. Second, if we need to find a phrase in the corpus in all its possible grammatical forms, we use the field 'Lexico-grammatical search'. In this case, in the column 'Word 1' we type the necessary word, while in the column 'Distance in words' set the necessary distance (for example, distance 1–1 means that the words will be adjacent to each other directly, distance 1–2 means there will be one word between the words, 1–3 means there is a distance of two words between the words, etc.), and in the column 'Word 2' we type the second necessary word. Thus, the corpus serves as a testing ground for verifying language choices in translation.

3 Results and Discussion

The examples to analyse were taken from the popular science books. The choice of the books was determined by the fact that we had a course in the popular science translation for master's students where we analysed different translations in this genre. Unlike in the belles-lettres, in popular science texts the language is far from being idiosyncratic. Rather, it is well standardised. This allows us to analyse the typical patterns of how the language is used by using the NCRL.

Let us analyse the first example.

The formulation of quantum theory made it necessary to rewrite James Clerk Maxwell's equations on electricity and magnetism. [1]/Formulirovka kvantovoy teorii zastavila perepisat uravneniya Dzheymsa Klerka Maksvella, kasayushchiyesya elektrichestva i magnetizma (The formulation of quantum theory forced the rewriting of James Clerk Maxwell's equations concerning electricity and magnetism) [20, p. 28].

Let us consider the translation correspondence *the formulation made it necessary to rewrite/formulirovka zastavila perepisat*. According to the explanatory dictionaries [7, 18, 30] and bilingual dictionaries [17], these correspondences are regular systematic cross-linguistic correspondences. The noun *formulation* is defined in the Oxford English Dictionary as: 1. mass noun The action of creating or preparing something. 1.1. count noun A particular expression of an idea, thought, or theory. 2. a material or mixture prepared according to a formula [18]. The Russian noun *formulirovka*, according to the Explanatory Russian Dictionary, is 1. only singular, Engagement in the formulation of the main provisions. 2. Formula, formulated position [30]. The English model verb *to make* is defined as: with object and infinitive Compel (someone) to do something [18], which is also a regular correspondence to the Russian model verb *zastavit*, i.e. to compel someone to do something

[7]. According to the bilingual, translational dictionaries, they are systematic cross-linguistic correspondences: *formulation/formulirovka*; *to make it necessary/zastavit*; *to rewrite/perepisat* [17].

Let us turn to the corpus data. The corpus data confirm that the linguistic form *formulirovka zastavila* is not natural in the Russian language. If you use the search for exact forms, then the search results are: nothing was found for this query. Let us use the lexico-grammatical search for these words in all its possible grammatical forms: in the field ‘Word’ we type the word *formulirovka*, while in the field ‘Word 2’ *zastavila* and set the distance 1–1; 1–2; 1–3. The results are: nothing was found for this query [15]. To sum it up, this linguistic form is not natural for the Russian language and is a speech error in translation, i.e. a collocation error. Importantly, if the author does not aim to achieve a stylistic effect, collocations errors are errors in using linguistic forms in speech [8].

Supposedly, if we produce the target text mostly through manipulations/transformations with the forms of the source text, the product of the process will exhibit an impact of the source language in the target text, i.e. it may result in some logical and/or linguistic/speech tension, or ‘stylistic discomfort’, as Yu. Sorokin put it [28]. This is so-called the ‘third language’ in translation, according to A. Duff [6], i.e. preserving the linguistic forms of the source text in the target text. In other words, preserving the language and speech norm of the source language during the second stage of translation semiosis, i.e. re-expressing the original meaning in the target language, may lead to collocation errors. Note that using the verb *zastavit* and the inanimate noun may be natural for the Russian language, yet this collocation can be found in fiction.

In the following example, there is also a collocation error in translation. However, if in the previous example the collocation error was largely due to preserving the forms of the source language in the target language, the collocation error in the following example is largely due to intralinguistic factors in the target language, i.e. blending seemingly similar phrases.

Newton formulated the basic description of how things move and effectively invented rocket science. [1]/Nyuton sformuliroval osnovnoye opisaniye dvizheniya obyektov i, v sushchnosti, izobrel printsip dvizheniya raket (Newton formulated the basic description of the movement of objects and, in essence, invented the principle of the movement of rockets) [20, p. 18].

Let us analyse the translation correspondence *formulated description/sformuliroval opisaniye*. According to the explanatory dictionaries [30] and bilingual dictionaries [17], these correspondences are regular systematic cross-linguistic correspondences. However, the search for exact forms in the NCRL (*sformuliroval opisaniye*) does not give any results as does not the lexical and grammatical search. It suggests that this linguistic form is not a natural, idiomatic, regular collocation in the Russian language. Assumingly, this collocation error in the Russian language is largely due to blending seemingly similar phrases: *sformulirovat polozheniya/printsipy i dat opisaniye (formulate provisions/principles and describe).*

Therefore, the question may arise: how can we avoid collocation errors in translation? In our opinion, a comparative study of the regular patterns of how linguistic forms are used in speech using the corpus methods is essential. This analysis will result in defining the frame and parameters to select possible translation variants [22]. This area of contrastive analysis represents the functional aspect of contrastive linguistics, which is of significantly greater importance for translation studies than the contrastive analysis of the language as a system. In other words, this is the difference between linguistic competence and linguistic performance, or language system and language use. If we approach translation as a speech production within the frame of activity type of translation ontology [10], the most effective tool to avoid speech errors in translation is a contrastive analysis of how the linguistic forms are used in speech by using the corpus-based methods [25].

Identifying the patterns of using linguistic forms in speech should result in constructing a lexico-semantic field of language forms of expressing the meaning in the source language and the target language [12]. To this end, we can use corpus-based methods, for example, the software Sketch Engine [27]. By using this software, we can build our own corpus on the given topic and register. It performs co-occurrence analysis, term extraction and generates frequency lists. Additionally, word sketches help identify idiomatically correct word combinations and help use words like native speakers do. It has a feature of topic modelling, i.e. keyword frequency, term extraction and term frequency will be useful for topic modelling by identifying words and phrases typical for the content of the text. In this regard, such lexico-semantic fields are sets of possible language choices in translation. Lexico-semantic fields represent language in its pre-speech condition, i.e. something between language system and language use. The construction of the lexico-semantic field of possible language choices is of particular importance at the second stage of translation semiosis, i.e., when the meaning given in the original, or translation invariant, should be re-expressed by linguistic forms of the target language. The focus is on a complex conscious choice of various possible translation solutions in the target language. The basis for identifying the language choices for expressing the meaning given in the source text in the target language is primarily the study of parallel, non-translational texts created in the target language on the subject of the source text and of the relevant functional style and register. Such a ‘translational’ study of parallel texts implies an identification of natural cross-linguistic correspondences. Knowing these typological features of the source and target languages in terms of how the linguistic forms are used in speech will optimise heuristic search in translation [19].

The integrative approach to contrastive analysis and translation studies using the corpus-based methods is widely discussed in modern Western studies [3, 5, 14]. Yet it has received little attention in Russia [24]. This underpins the novelty of this article. Integrating contrastive analysis, translation studies, and corpus-based methods is important, since integrative approach can significantly enrich both translation studies and contrastive analysis with new approaches to understanding fundamental categories: translation equivalence, translation norms, translation quality assessment, translation invariant, etc. Additionally, on the level of practice, the results of the study can be used to improve the criteria for the translation quality assessment.

4 Conclusion

To sum up, the article presented an analysis and verification of language choices in translation in terms of the communicative equivalence using corpus methods. The analysis showed that translation solutions can be correspondences at the level of the language system, yet they are not pragmatically equivalent solutions. In other words, in translation there are cross-linguistic correspondences that are identified at the level of the language as a system, without taking into account the regular patterns of how they are naturally used in speech. This results in speech errors, i.e. mostly, collocation errors. In this regard, incorporating corpus-based methods into contrastive analysis for translation purposes is essential.

Such corpus studies that focus on identifying the regular patterns of how linguistic forms are used in speech are undoubtedly of heuristic value not only for the theory and practice of translation, but also for translation quality assessment. The contrastive analysis that focuses on language use, rather than language system, in contrast is essential for translation. This became possible only with the advent of corpus-based methods and Internet technologies.

References

1. 30-second Theories: The 50 most thought-provoking theories in science. Ed.: Paul Parsons. Icon, London (2011).
2. Alekseeva, I. S. (2004). *Vvedeniye v perevodoyevedeniye (Introduction to translation studies) in Russian Academy.*
3. Altenberg, B., & Granger, S. (2002). *Lexis in contrast. Corpus-based approaches.* John Benjamin Publishing Company.
4. Barkhudarov, L. S. (2021). *Yazyk i perevod (Language and translation) in Russian.* Librokom.
5. Doval, I., Teresa Sanchez Nieto, M. (2019) Parallel corpora for contrastive and translation studies. New resources and applications. John Benjamin Publishing Company.
6. Duff, A. (1981). *The third language: Recurrent problems of translation.* Pergamon Institute of English Press.
7. Efremova, T.F. (2022) Novyy slovar russkogo yazyka. Tolkovo-slovoobrazovatelnyy. (New dictionary of the Russian language. Explanatory derivational) in Russian. Last accessed May 1, 2022, http://www.efremova.info/.
8. Golub, I. B. (2007). *Stilistika russkogo yazyka (Stylistics of the Russian language) in Russian.* Iris-Press.
9. Kazakova T.A. (2006) Khudozhestvennyy perevod. Teoriya i praktika (Literary translation. Theory and practice) in Russian. Inyazizdat
10. Kryukov, A. N. (1989). *Teoriya perevoda (Translation theory) in Russian.* Publishing House of the Military Institute.
11. Lekomtseva I.A., Kuraleva T.V. (2018) Mezhyazykovaya asimmetriya pri perevode (Interlingual asymmetry in translation) in Russian. *Baltic Humanitarian Journal* 7(1(22)), 101–105.
12. Lekomtseva I.A., Vyunova E.K., Abdulmanova A.Kh. (2022) Opredeleniye vozmozhnykh pere-vodcheskikh resheniy v ramkakh kontseptsii yestestvennoy ekvivalentnosti na osnove sopostavitelnykh issledovaniy s pomoshchyu korpusnykh metodov (Determination of possible translation solutions within the framework of the concept of natural equivalence based on comparative studies using corpus-based methods) in Russian. Nauchnii Dialog (in print).

13. Leontiev, A. A. (2003). *Yazyk, rech, rechevaya deyatelnost (Language, speech, speech production) in Russian.* URSS.
14. Mikhailov, M., & Cooper, R. (2016). *Corpus linguistics for translation and contrastive studies (Routledge corpus linguistics guides).* Routledge.
15. National corpus of the Russian language. Last accessed May 1 2022, https://ruscorpora.ru/.
16. Nauchnyye teorii za 30 sekund. 50 samykh genialnykh nauchnykh teoriy, rasskazannykh za polminuty (Scientific theories in 30 seconds. 50 most ingenious scientific theories told in half a minute) in Russian. Per. from English. Y. Kapustyuk; scientific ed. A. Nurmatov; ed. Paul Parsons. Moscow: RIPOL Classic, 2013.
17. Online bilingual dictionary system Multitran. Last accessed May 1 2022, https://www.multitran.com/.
18. Oxford English dictionary, Last accessed May 1 2022, https://lexico.com.
19. Petrova E.S. (2011) Sopostavitelnaya tipologiya angliyskogo i russkogo yazykov. Grammatika (Comparative typology of English and Russian languages. Grammar) in Russian. Faculty of Philology, St Petersburg State University, St Petersburg; Academy.
20. Psikhologiya za 30 sekund (2013) 50 samykh genialnykh teoriy v psikhologii, kazhdaya iz kotorykh obyasnyayetsya rasskazannykh za polminuty (Psychology in 30 Seconds: 50 of the most ingenious theories in psychology, each of which is explained in half a minute) in Russian. Per. from English. Yu. Zmeeva; scientific ed. A. Nurmatov; ed. Christian Jarrett. Moscow: RIPOL Classic.
21. Pym, A. (2017). *Exploring translation theories.* Routledge.
22. Retsker Ya.I. (2010) Teoriya perevoda i perevodcheskaya praktika: ocherki lingvisticheskoy teorii perevoda (Translation theory and translation practice: essays on the linguistic theory of translation) in Russian. R. Valent, Moscow.
23. Riabtseva N. K. Perevodovedeniye v Rossii i za rubezhom. Ch. 2. Analiz empiricheskogo materiala (Translation Studies in Russia and Abroad. Part 2. A case study) in Russian. last accessed May 1, 2022 https://iling-ran.ru/riabtseva/translationstudies2.pdf.
24. Ryzhova D.A. (2020) Tipologiya leksiki. Kompyuternyye metody i instrumenty (Typology of lexis. Computer methods and tools) in Russian. Alateya, St. Petersburg.
25. Schweitzer A.D. (2020) Kontrastnaya stilistika. Gazetno-publitsisticheskiy stil v angliyskom i russkom yazykakh (Contrasting stylistics. Newspaper-journalistic style in English and Russian) in Russian. Librokom, Moscow.
26. Shcherba, L. V. (2008). *Yazykovaya sistema i rechevaya deyatelnost (Language system and speech production) in Russian.* LKI.
27. Sketch engine, last accessed January 21, 2022, http://sketchengine.eu.
28. Sorokin, Yu. A. (2003). *Perevodovedeniye: Status perevodchika i psikhogermenevticheskiye protsedury (Translation studies: The status of a translator and psychohermeneutic procedures) in Russian.* Gnosis.
29. Sternin I.A. (2007) Kontrastnaya lingvistika. Problemy teorii i metodiki issledovaniya (Contrasting linguistics. Problems of theory and methods of research) in Russian. AST.
30. Ushakov D.N. Tolkovyy slovar russkogo yazyka (Explanatory dictionary of the Russian language). Last accessed May 01, 2022 (in Russian), http://ushakovdictionary.ru/.
31. Vygotsky, L.S. (2021) Myshleniye i rech (Mind and speech) in Russian. Publishing house ‘Piter’.

Corpora in…… Literature Studies

Stylometric Methods in Comparative Text Analysis

Alexander Grebennikov, Ekaterina Ivanova, Mikhail Koryshev, and Maria Solovieva

Abstract The article focuses on the study of the possible objectification of the procedure for comparative analysis of the original text and translation texts using computer technologies. The study draws on a lecture by Vladimir Nabokov on the novel "Madame Bovary" by Gustave Flaubert in Russian, English, French, and German. Using the T-LAB computer program, frequency dictionaries were built for the analyzed texts, which made it possible to identify ten words from the upper zone of the obtained frequency distributions that best correspond to the main themes of the novel. In addition, with the help of the T-Lab program, the associative links of the selected lexemes were analyzed. Based on the association analysis results, a number of coincidences of the joint occurrence of the analyzed lemma with other lemmas for pairs of languages were calculated. It has been established that the frequency of selected keywords demonstrates a significant match in all the languages, while in some cases there is a translation of an English word with the help of two words in the target language. However, at the level of contextual association the number of matches decreases, nowhere reaching a 100% match. Therefore, the applied method of computer stylometric analysis makes it possible to trace how the translator, creatively interpreting the meaning of the source text, forms their own associative fields to reveal the content of the author's topics.

Keywords Nabokov · Stylometry · Frequency dictionary · Translation studies · Russian literature

A. Grebennikov (✉) · E. Ivanova · M. Koryshev · M. Solovieva
St. Petersburg University, Universitetskaya Emb. 7/9, 199034 St. Petersburg, Russia
e-mail: a.grebennikov@spbu.ru

E. Ivanova
e-mail: eivanova2003@mail.ru

M. Koryshev
e-mail: mkorychev@yandex.ru

M. Solovieva
e-mail: mvsol77@mail.ru

P. Eismont et al. (eds.), *Literature, Language and Computing*,
https://doi.org/10.1007/978-981-99-3604-5_9

1 Introduction

Over the past decades, the methodology of linguistic research has undergone dramatic changes as a result of the formation and development of new methods of language analysis and methods of its description, in particular, the use of computer technology. Thus, the methods of automatic language processing in its written and oral forms turned out to be in demand in the field of corpus and lexicographic research, text attribution, literary text analysis, stylometry, modeling of language processes, language forecasting, etc. [1, 2].

A different situation is observed in the field of translation studies, in which the use of computer tools, including information technologies for data processing, is mainly aimed at optimizing the translation process in professional translation activities (using translation memory systems and machine translation). Thus, the potential resource of computer technologies, limited by a specific applied task, remains out of the focus of researchers' attention in their theoretical understanding of the fundamental concepts of translation studies (equivalence and adequacy).

The study aims to fill this gap and specify how computer technologies make it possible to objectify the procedure for a comparative analysis of the original and translated texts. For the first time in the research of the kind, we propose a technique validating the analysis of the associative links of selected lexemes.

In the framework of the study, the authors used a licensed version of the T-LAB computer program of the Italian manufacturer [3], which includes a set of linguistic, statistical and graphical tools for various types of text analysis (content analysis, sentiment analysis, semantic analysis, thematic analysis, text mining, perception mapping, discourse analysis, and network analysis). The corpus imported into the program can be presented both as a separate text and as a set of texts, while the latter option seems the most promising for conducting a comparative study of multilingual corpora.

An extract from Vladimir Nabokov's lectures known as 'Lectures on Literature' [4], focusing on the analysis of the novel 'Madame Bovary' by Gustave Flaubert, served the basis of the parallel multilingual corpus of this study. The lectures were delivered by V.V. Nabokov at Wellesley College and Cornell University during the American period of his career in the 1940s–50 s and addressed the works by Russian (Gogol, Turgenev, Dostoevsky, Tolstoy, Gorky) and foreign writers (Austen, Dickens, Flaubert, Joyce, Kafka, Proust, Stevenson and Cervantes). For a comparative analysis, translations of Nabokov's lecture on G. Flaubert into Russian [5], French [6], and German [7] were used. It is worth noting that the posthumous publication of V.V. Nabokov's lectures, which the authors of this article owe primarily to his wife Vera Nabokov, was acknowledged as a momentous event both by modern literary critics and Nabokov scholars. Despite some critical remarks on the part of a few literary critics after the book was published, it should be noted that Nabokov's text is not so much a canonical example of the analysis of a literary work but rather a literary text in itself. Through the prism of this text V.V. Nabokov presents himself as a writer-reader [8].

2 Literature Review

Since the second half of the last century, the discussion of equivalence has been in the focus of translation studies [9], which is evidenced, in particular, by the name of the specialized journal on translation theory “Equivalences”, published by the Graduate School of Translation in Brussels [10]. The concept of equivalence, borrowed from logic, has to a certain extent replaced unclear and not particularly objective normative criteria of ‘fidelity’, ‘accuracy’ and ‘correspondence’. With the development of the theory of translation, as well as linguistic research discourse, this concept has been transformed in accordance with new ideas [11–14]. At the same time, despite a significant number of definitions of this concept and many attempts to define it, the fact remains that equivalence is, first of all, a measure of the accuracy of a translation in relation to the original.

Despite the widespread opinion that a large number of definitions make it easier to develop the right approach within the framework of a specific research problem, it is difficult to agree with this point of view. The lack of the unified definition, as well as clear characteristics of a particular phenomenon, seriously complicates further research. Currently researchers almost unanimously agree on only one thing: when studying such a concept as “equivalence”, it is necessary to take into account the multilayer nature of this phenomenon, be it multilevel hierarchical systems [15–18] or continual fragments limited by the two poles, i.e. formal correspondence and dynamic equivalence [19]. The latter one determines the degree of the effect that the translated text has on the recipient, as opposed to the effect that the source text has on the recipient of the original, and between the two poles—dynamic equivalence and formal correspondence—there are many intermediate options. This fact seems to be especially important in relation to the translation quality assessment. Indeed, to measure the extent of equivalence more precisely, and rather the degree of deviation from the principles of equivalence, as well as a good reason behind such deviation, is the main task in assessing the quality of a translation.

With time, researchers refused to base the entire theory of translation on the concept of equivalence, saying that it is only one of the many factors for successful translation, but not the major one. So, according to the proponents of functional linguistics, equivalence is only one of the many goals of translation which can be achieved or not, depending on the skopos that is set for the translation text as a whole [20–22].

Disputes about the theoretical status of the concept of equivalence in translation studies do not subside even today. Thus, a “pragmatic approach” is being brought forward to understand the equivalence and the “loyaute” of translation [23], in which equivalence is defined as the requirement for the obligatory transfer of the pragmatic textual meaning of the source text in the translation. Also, K. Nord’s idea about the need to take into account the type of text (documentary vs instrumental) is becoming more widespread. In both cases, these requirements relate only to the translation as a product (the end result of the translation process), which in this case turns out to be a normative criterion for assessing the quality of a translation. It should be noted that in modern translation studies the concepts of equivalence and normativity, assessment and parametrization of translation are inevitably connected. In other words, the evaluative nature of the category is universally recognized, and, therefore the question of how the degree of equivalence can be measured is attracting more and more attention of the researchers. However, the academic papers known to us in this field are, as a rule, limited to didactic purposes and are aimed at creating classifications of errors/mistakes and text evaluation scales for the translation of documentary texts. At the same time, the translation of literary texts from this standpoint has not been studied before.

An important aspect of studying the stylistic features of a particular work is the analysis of its lexical originality: it is the vocabulary that forms what linguists call the language picture of the world. When considering a set of texts, the frequency distribution of words is of great importance, and this applies primarily to meaningful parts of speech (nouns, adjectives, numerals, verbs, adverbs). Word frequency analysis has proved instrumental in language studies, in general, and in corpus linguistics, in particular [24, 25]. Apart from this, this kind of analysis may be helpful in bringing out lexical features characteristic of a writer's individual style [26]. Moreover, it has been shown to allow detecting dynamics in the thematic content of the texts under investigation [27–29].

3 Methodology

The texts selected for analysis were processed with the help of the T-LAB program. The program opens up a wide range of opportunities for the research analysis based on the frequency dictionaries of the studied texts created by it.

At the initial stage, with the help of the program, frequency dictionaries of lexemes of the analyzed texts were built (English text—19 386 tokens, 3 969 lexemes; German text—18 731 tokens, 4 592 lexemes; French text—19 977 tokens, 3 934 lexemes; Russian text—14 558 tokens, 4 762 lexemes). When a dictionary is being built up, polysemantic words are most often considered in the aggregate of all their variants, which, of course, leads to certain errors, which is undoubtedly an inevitable consequence of the use of automatic methods.

At the next stage, the following key words (themes) were identified by means of textual analysis in the text of the lecture, i.e. the names of the characters—Emma and Charles; the key themes of the novel, i.e. *love, death* and the word *theme*; words characterizing the style of the work as a literary device, i.e. *horse* as the tenor of the novel, *counterpoint* as a literary device, *romantic* as a definition of the style of the literary work. And, finally, *philistine/bourgeois* as a definition of the socio-historical context of the novel.

As a pilot study, the authors chose ten words from the top zone of the obtained frequency distributions that best correspond to these themes.

Strictly speaking, there are more words than were originally chosen, since the word *romantic* is represented by two lexemes in the French text; also the difference can be witnessed in rendering of mutually related concepts in the original text of the *bourgeois* in different languages. Lexemes, their frequencies, as well as ranks (static weights) in the analyzed texts are shown in Tables 1 and 2.

In general, the above data demonstrate, for the most part, a high similarity of the frequencies and/or ranks of the selected lexemes, sometimes with a literal match. It is worth noting that we are talking about texts that are to a large extent identical in content (the same original text translated in three languages). Therefore, we do not manipulate here with such traditional concept of linguistic statistics as sampling and all the tools associated with it; and where there is a similarity in frequencies and/or ranks, or, on the contrary, there is their significant difference, the authors reasonably associate it with the general features of the translation.

Based on the obtained dictionaries, the T-Lab program allows us to analyze the associative links of the selected lexemes. Joint occurrences of the analyzed lexeme are calculated within its elementary contexts. At the next stage, the number of coincidences of the joint occurrences of other lemmas with the analyzed lemma was calculated for pairs of the languages (English–Russian; French–Russian; German–Russian). The results are presented in Tables 3, 4 and 5.

Table 1 Comparison of ranks and frequencies of the selected key lexemes in English and German texts

English			German		
Rank	Lexeme	Frequency	Rank	Lexeme	Frequency
1	Emma	160	1	Emma	166
2	Charles	72	2	Charles	71
23	Horse	22	24	Pferd	22
11	Love	36	36	Liebe	20
55	Death	15	79	Tod	14
28	Romantic	19	110	Romantischen	12
78	Style	12	54	Stil	12
26	theme	21	176	Motiv	10
				Thema	8
122	Counterpoint	9	181	Kontrapunktisch	5
				Kontrapunkt	4
44	Philistine	16	133	Spießer	10
75	Bourgeois	12	243	Bürger	6
			247	Bourgeois	5

Table 2 Comparison of ranks and frequencies of the selected key lexemes in French and Russian texts

French			Russian		
Rank	Lexeme	Frequency	Rank	Lexeme	Frequency
1	Emma	134	1	Emma	165
3	Charles	57	2	Sharl	64
20	cheval	25	15	loshad’	20
25	Amour	21	17	lubov’	20
42	Mort	16	37	smert’	14
41	Romanesque	16	20	romantisheskiy	18
259	Romantique	6	–	–	–
65	Style	13	41	stil’	13
29	Thème	19	21	tema	18
147	Contrepoint	8	59	kontrapunkt	11
50	Bourgeois	15	70	meshchanin	10
171	Philistin	8	308	burzhua	5
			357	bourgeois	4

Table 3 Comparison of the number of joint occurrences of other lemmas with the analyzed lemma in English and Russian texts

English			Russian			Number of joint occurrences (out of 20)
Rank	Lexeme	Frequency	Rank	Lexeme	Frequency	
1	Emma	160	1	Emma	165	9
2	Charles	72	2	Sharl	64	6
23	Horse	22	15	loshad’	20	4
11	Love	36	17	lubov’	20	2
55	Death	15	37	smert’	14	1
28	Romantic	19	20	romantisheskiy	18	7
78	Style	12	41	stil’	13	5
26	Theme	21	21	tema	18	7
122	Counterpoint	9	59	kontrapunkt	11	7
44	Philistine	16	70	meshchanin	10	Jointly 12
75	Bourgeois	12	308	burzhua	5	
			357	bourgeois	4	

Table 4 Comparison of the number of joint occurrences of other lemmas with the analyzed lemma in French and Russian texts

French			Russian			Number of joint occurrences (out of 20)
Rank	Lexeme	Frequency	Rank	Lexeme	Frequency	
1	Emma	134	1	Emma	165	12
3	Charles	57	2	Sharl	64	6
20	cheval	25	15	loshad’	20	4
25	amour	21	17	lubov’	20	3
42	mort	16	37	smert’	14	3
41	romanesque	16	20	romantisheskiy	18	6
259	romantique	6				
65	style	13	41	stil’	13	8
29	thème	19	21	tema	18	5
147	contrepoint	8	59	kontrapunkt	11	7
50	bourgeois	15	70	meshchanin	10	Jointly 8
171	philistin	8	308	burzhua	5	
			357	bourgeois	4	

Table 5 Comparison of the number of joint occurrences of other lemmas with the analyzed lemma in German and Russian texts

German			Russian			Number of joint occurrences (out of 20)
Rank	Lexeme	Frequency	Rank	Lexeme	Frequency	
1	Emma	166	1	Emma	165	8
2	Charles	71	2	Sharl	64	7
24	Pferd	22	24	loshad’	20	8
36	Liebe	16	36	lubov’	20	2
79	Tod	9	79	smert’	14	5
110	romantischen	8	110	romantisheskiy	18	4
54	Stil	12	54	stil’	13	4
176	Motiv	6	176	tema	18	8
181	kontrapunktisch	6	181	kontrapunkt	11	3
133	Spießer	7	133	meshchanin	10	Jointly 14
243	Bürger	5	243	burzhua	5	
247	Bourgeois	5	247	bourgeois	4	

4 Results

See Tables 1, 2, 3, 4 and 5.

5 Discussion

As follows from the presented results, the frequency of selected keywords (Tables 1 and 2) reveals a significant match, while in some cases there is a translation of an English word with two words in the target language (e.g., translation of *romantic* by two French lexemes *romanesque/romantique*, English *theme* with German *motiv/thema*, *counterpoint* with German *kontrapunktisch/kontrapunkt*). There is also a difference in rendering of the concepts *philistine* and *bourgeois*, mutually related in the original text, in different languages. However, when we proceed to the level of contextual association (Tables 3, 4 and 5), the number of matches decreases, nowhere reaching 100% match. The maximum number of coincidences in associations is observed in the names of characters (at 60%), which seems quite natural given the thematic focus of the text. With regard to translations of other keywords used for text analysis, there is a decreasing tendency for the proportion of matches in the translation of the author's words (*counterpoint, style, theme, horse*, etc.) to words denoting general concepts (*love, death*), for which the proportion coincidence is sometimes only 15%. At the same time, some variability in the proportions of correspondences cannot but be observed, which, in our opinion, is naturally due to the difference in the systems of the languages in question. Therefore, the translator, creatively interpreting the meaning of the source text, shapes his own associative fields to reveal the content of Nabokov's themes.

6 Conclusion

It has been established that the use of the T-LAB computer program which includes a wide range of stylistic tools for various types of text analysis, allows for an objective comparison of the original text and its translations into various languages. In addition to confirming the ability of such analysis to identify keywords that reflect the main themes of the text established in previous studies [27–29], a wide range of opportunities to analyze the associative links of selected lexemes by computer methods have been demonstrated. The study has demonstrated that the analysis of correspondences of frequencies and associations from language to language makes it possible to trace the process of the translator's creation of their own associative fields when they work with the author's text. Since the above observation brings the authors to the concept of the equivalence of the translated text as related to the source text, it seems reasonable at the next stage to turn to an in-depth analysis of the "non-coincident" associative links of lexemes in translation and to the ways that make it possible to achieve equivalence depending on the functional nature of the text. Furthermore, the results obtained may be useful on the theoretical level, when discussing the notion of equivalence itself from a new perspective.

Acknowledgements This research has been conducted with financial support from St. Petersburg University (project No 92565342).

References

1. Martynenko, G. (1998). *Osnovy stilemetrii [Basics of stylemetria]*. Izdatel'stvo Sankt Peterburgskogo universiteta. (In Russian).
2. Lasswell, H. D. (1965). Style in the language of politics. Language of Politics: Studies in Quantitative Semantics. MA, M.I.T. Press, Cambridge, 20–39.
3. T-LAB, https://www.tlab.it/, last accessed 2021/09/10.
4. Nabokov, V. (1980) Nabokov Vladimir. Lectures on Literature. Introduction by John Updike. Edited by Fredson Bowers, New York City, 125–178.
5. Nabokov, V.V. (1998). Nabokov V.V. Lekcii po zarubezhnoj literature [Lectures on foreign literature]. Izdatel'stvo nezavisimaya gazeta, Moskva, 183–238 (In Russian).
6. Nabokov, V. (2009). Nabokov Vladimir. Littératures. Introductions de John Updike et de Guy Davenport. Editions Robert Laffont, Paris, 191–250.
7. Nabokov, V. (1982). Die Kunst des Lesens: Meisterwerke der europäischen Literatur. Vorw. von John Updike. S. Fischer, Frf. am Main, 173–229.
8. Guseinova, E.R. (2010). Igrovye strategii v lekcionnom kurse V. Nabokova [Games in Navokov's lecture course]. Izvestiya Saratovskogo universiteta 10(2), 74–77 (In Russian).
9. Prunč, E. (2007). *Entwicklungslinien der Translationswissenschaft: Von den Asymmetrien der Sprachen zu den Asymmetrien der Macht*. Frank & Timme.
10. Equivalences: revue de traduction et de traductologie, http://www.trad–equivalences.org, last accessed 2022/12/04.
11. Bolanos Cuéllar, S. (2002). Equivalence Revisted: A key Concept in Modern Translation Theory. *Forma y Functión, 15*, 60–88.
12. Draskau, J. K. (1991). Some Reflections on "Equivalence/Aquivalenz" as a Term and Concept in the Theory of Translation. *Meta, 36*(1), 269–274.
13. Melby, A. (1990). The Mentions of Equivalence in Translation. *Meta, 35*(1), 207–213.
14. Pym, A. (2010). Translation Theory Today and Tomorrow—Responses to Equivalence. In L. N. Zybatow (Ed.), *Ranslationswissenschaft—Stand und Perspektiven* (pp. 1–14). Frankfurt aM.
15. Jäger, G. Translation und Translationslinguistik. Veb Max Niemeyer, Halle.
16. Koller, W. (1979). *Einfuhrung in die Übersetzungswissenschaft*. Quelle & Meyer.
17. Komissarov, V. (1987). The semantic and the cognitive in the text: A problem in equivalence. *Meta, 32*(4), 416–419.
18. Neubert, A. (2000). Competence in language, in languages, and in translation. In: C. Schäffner, B. Adab (Eds.), Developing translation competence. John Benjamins Pub, Amsterdam–Philadelphia, 3–18.
19. Nida, E.A. (1964). Toward a science of translating. Brill, Leiden.
20. Reiss K. (2002). La critique des traductions. Ses possibilités et ses limites: cathégories et critères pour une évaluation pertinente des traductions. Artois Presses Université, Artois.
21. Vermeer, H.J. (2000). Skopos and Commission in Translational Action (translated By A. Chesterman). The Translator Studies Reader. Routledge, London and New York, 221–232.
22. Nord, Ch. (1996). Translating as a purposeful activity. Functionalist approaches explained. St. Jerome, Manchester.

23. Emery, P. G. (2004). Translation, equivalence and fidelity: a pragmatic approach. *Babel, 50*(2), 143–167.
24. Leech, G., Rayson, P., & Wilson, A. (2001). *Word Frequencies in written and Spoken English: Based on the British National Corpus*. Longman.
25. Baron, A., Rayson, P., Archer, D. (2009). Word frequency and key word statistics in historical corpus linguistics. *Anglistik: International Journal of English Studies* 20(1), 41–67.
26. Grebennikov, A., & Skrebtsova, T. (2019). Yazykovaya kartina mira v russkom rasskaze nachala XX veka [Language world picture in Russian stories of the beginning of XX century]. *Philosophy and the Humanities in the Information Society, 3*, 82–92. (In Russian).
27. Sherstinova, T., Grebennikov, A., Skrebtsova, T., Guseva, A., Gukasian, M., Egoshina, I., Turygina, M. (2020). Frequency Word Lists and Their Variability (the Case of Russian Fiction in 1900–1930). In: *27th Conference of Open Innovations Association FRUCT, University of Trento, Italy*, 366—373, https://fruct.org/publications/acm27/files/She.pdf.
28. Sherstinova, T., Skrebtsova, T. (2021). Russian Literature Around the October Revolution: A Quantitative Exploratory Study of Literary Themes and Narrative Structure in Russian Short Stories of 1900–1930. In: Proceedings of the International Workshop «Computational Linguistics» (St. Petersburg, 17–20 June, 2020), CEUR Workshop Proceedings, 2813, pp. 117—128, http://ceur–ws.org/Vol–2813/rpaper09.pdf.
29. Skrebtsova, T. (2021). Thematic tagging of literary fiction: the case of early 20th century Russian short stories. In: *Proceedings. of the International Workshop* «Computational Linguistics» (St. Petersburg, 17–20 June, 2020), CEUR Workshop Proceedings, 2813, pp. 265—276, http://ceur–ws.org/Vol–2813/rpaper20.pdf.

Lexical Diversity of Russian Poets

Alexander Piperski

Abstract The paper discusses the lexical diversity of the works by 183 Russian poets included in the Poetic subcorpus of the Russian National Corpus. The study employs seven measures of lexical diversity (five versions of Standardised Type-Token Ration, Simpson's diversity index, and Zipf–Mandelbrot's *a*), which are linked to different aspects of interpretation, namely preference to / avoidance of lexical repetition, topical homo-/heterogeneity, and overall complexity of language and style, including vocabulary richness. A corpus-driven approach makes it possible to analyze how lexical diversity is related to two sociolinguistic variable present in the markup of the Poetic subcorpus, namely birth year and gender. It is demonstrated that topical heterogeneity and complexity of language and style have been on the rise since the eighteenth century. Male poets exhibit larger values of diversity linked to topic heterogeneity, but this only manifests itself in the use of nouns and verbs and does not apply to adjectives.

Keywords Lexical diversity · Corpus linguistics · Russian poetry · Gender · Parts of speech

1 Introduction

The study of lexical diversity has a long-rooted tradition in stylistics, corpus linguistics, and L2 acquisition [1–3]. The advent of large annotated corpora has made it possible to study and compare lexical diversity of different authors, time periods, etc. The aim of this paper is to perform an analysis of lexical diversity in Russian poetic texts from the last three centuries.

A. Piperski (✉)
Russian State University for the Humanities, 15 Chayanova Str, 125047 Moscow, Russia
e-mail: apiperski@gmail.com

P. Eismont et al. (eds.), *Literature, Language and Computing*,
https://doi.org/10.1007/978-981-99-3604-5_10

It is hard to estimate lexical diversity impartially because the assessment of lexical diversity often involves value judgments. Smaller lexical diversity has been linked to undesirable medical conditions, such as Alzheimer's disease [4] or dementia [5]. Lexical diversity is known under various names and is often mentioned together or even confused with lexical richness, which is the number of lexical items an individual knows [6], and it is often implied that a larger vocabulary is better than a smaller one. Be that as it may, it should be kept in mind that lexical diversity is a multifaceted phenomenon, and it is not always the case that a text that exhibits more lexical diversity is of better quality than a text that has smaller diversity. For instance, one would hardly question the artistic quality of Homeric poems; however, it is well-known that they exhibit a large degree of repetition [7], which has a negative impact on their lexical diversity. On the other hand, a text with the highest possible value of lexical diversity is a spelling dictionary, in which no item is repeated twice; however, no one would probably ascribe high artistic or stylistic merit to such a text.

Any index of lexical diversity subsumes a number of various properties of a text, its author, and its language. Low diversity may be an indicator of the author's inclination to use lexical repetition, of topical homogeneity of the text, of the author's preference for simpler and more understandable language, and of the author's vocabulary size. When comparing texts from different time periods, one should be also aware that a difference in lexical diversity may be caused by differences in grammar; e.g., any text in Modern English is likely to be deemed less diverse that a corresponding text in Old English, since it abounds in articles *a* and *the*.

Lexical diversity can be regarded as an important characteristic of an author. The present study, which focuses on lexical diversity of Russian poets, was performed using a corpus-driven rather than a corpus-based approach [8]. Its aim is not to support any pre-existing hypotheses concerning differences in lexical diversity between individual authors or groups of authors, but rather to identify patterns that emerge from the data which may be further subjected to more robust statistical testing and interpretation.

2 Materials and Methods

The study of lexical diversity of Russian poets was conducted using the texts included in the Poetic subcorpus of the Russian National Corpus (RNC). As of 2022, the Poetic subcorpus contains approximately 13 million tokens by 973 authors. However, for many of these poets the number of texts in the corpus is too small (in some cases it is only one short poem), and for this reason only 183 poets represented by more than 10,000 tokens of lexical words (nouns, verbs, and adjectives) were included in the sample. It would probably not be an exaggeration to say that this sample includes all well-known Russian poets and a large number of second-tier poets. For each author in the sample, date of birth and gender is available. All poems by one author are ordered by creation date and treated as a single text. The texts were lemmatized and disambiguated using DeepPavlov Python library [9]; all subsequent analyses were performed using lemmas tagged as nouns, adjectives, and verbs. It should be noted that Russian poetic texts are notoriously challenging for morphological processing [10]; however, one can probably assume that errors made by the lemmatizer do not have a strong impact on the calculation of lexical diversity.

No single measure of lexical diversity has proven to be applicable in all cases. One of the crucial problems with measuring lexical diversity is that we need a measure that is comparable for texts of various length [11, 12], which is of great importance for the sample of texts used in this paper, the longest of them (the poems by Vasily Zhukovsky) containing 171,084 lexical words and being 17 times longer than the shortest one (Pavel Zaltsman, 10,099 lexical words). For the purposes of the present study, I compute three measures of lexical diversity, the first of them coming in five versions:

1. Standardized Type-Token Ratio for spans of different length, denoted as STTR(N) [13], i.e. the mean number of different tokens in a moving window of N tokens ($N = 25$, 100, 500, 1000, and 5000).
2. Simpson diversity index λ (also known as Herfindahl or Hefindahl–Hirschmann diversity index):

$$\lambda = \sum_{i=1}^{V} p_i^2,$$

where V is vocabulary size and p_i is the relative frequency of the i-th vocabulary item. λ is the probability that two tokens randomly drawn from the text belong to the same type. Though rarely used as a measure of lexical diversity [14], which is probably due to the fact that it treats texts as unordered bags of words, this index has been proven to be very robust with respect to corpus size [11]; its bag-of-word nature is actually an advantage for the present study because it is not influenced by the order of poems being concatenated into a single text. The inverse of λ is the effective number of types observed in a text, i.e. how many same-frequency types a text with the same λ would contain.

3. Zipf–Mandelbrot's a, which is the exponent of the denominator in the formula for Zipf–Mandelbrot's law:

$$f = \frac{C}{(r + b)^a},$$

where f is the frequency of the word with the rank r, and a, b, and C are parameters estimated using the least-squares method. This measure is once again insensitive to the order of the poems by a single author.

The interpretation of these measures is far from straightforward. However, their definitions make it possible to establish at least tentative links to the aspects of lexical diversity identified in the Introduction:

1. STTR with small values of N reflects the inclination to / avoidance of lexical repetition, whereas STTR with large values of N reflects topical homo-/ heterogeneity.
2. Small values of λ (i.e., low probability of sampling two identical tokens from the text) hint at topical heterogeneity, whereas large values of λ (i.e., high probability of sampling two identical tokens) hint at topical homogeneity.
3. Small values of a (i.e., a gentle decrease in frequency from the top of the frequency list to its bottom) indicate lexical richness and a tendency to use more complex language, whereas large values of a (i.e., a steep decrease in frequency) indicate a preference for simpler language.
4. For all 183 poets, all seven measures were calculated; for 123 poets whose texts primarily stem from the twentieth century, an extra analysis involving the calculation of λ by different parts of speech was performed.

3 Results

Table 1 lists five most diverse and five least diverse poets according to each of the seven measures used.

Rank	STTR (5000)	Λ	a
1	Tsvetkov	Kenjeev	Narbut
2	Klyuev	Narbut	Parshchikov
3	Kenjeev	Klyuev	Tarlovsky
4	Narbut	Tsvetkov	Tsvetkov
5	V. Rozhdestvensky	Burliuk	Klyuev
…			
179	Bozhnev	Bozhnev	A. Grigoryev
180	Chemnitzer	Kharms	Cherubina de Gabriak
181	Kharms	Kropivnitsky	Baltrušaitis
182	Levitansky	Levitansky	Karamzin
183	Kropivnitsky	Chemnitzer	Chemnitzer

Most of these names probably do not tell much to a reader who is not an expert in Russian poetry. However, even a quick look at the names being repeated multiple times shows that the seven measures are intercorrelated. This is further illustrated in Table 2, where pairwise correlations between the measures are given; most pairs exhibit strong correlations, and only the correlation between STTR(25) and a is quite week.

Table 1 Five most diverse and five least diverse poets according to different measures

Rank	STTR (25)	STTR (100)	STTR (500)	STTR (1000)
1	Tsvetkov	Kenjeev	Narbut	Narbut
2	Klyuev	Narbut	Parshchikov	Tarlovsky
3	Kenjeev	Klyuev	Tarlovsky	Parshchikov
4	Narbut	Tsvetkov	Tsvetkov	Obolduev
5	V. Rozhdestvensky	Burliuk	Klyuev	Tsvetkov
…				
179	Bozhnev	Bozhnev	A. Grigoryev	Cherubina de Gabriak
180	Chemnitzer	Kharms	Cherubina de Gabriak	Karamzin
181	Kharms	Kropivnitsky	Baltrušaitis	Khomyakov
182	Levitansky	Levitansky	Karamzin	Baltrušaitis
183	Kropivnitsky	Chemnitzer	Chemnitzer	Chemnitzer

Table 2 Pairwise correlations (Pearson's *r*) between the seven lexical diversity measures[1]

	STTR (100)	STTR (500)	STTR (1000)	STTR (5000)	λ	*a*
STTR (25)	0.91	0.60	0.45	0.23	−0.23	−0.08
STTR (100)		0.86	0.75	0.54	−0.51	−0.25
STTR (500)			0.98	0.88	−0.81	−0.47
STTR (1000)				0.95	−0.86	−0.54
STTR (5000)					−0.86	−0.59
λ						0.31

Table 3 Correlation between birth year and lexical diversity: Pearson's *r*

	STTR (25)	STTR (100)	STTR (500)	STTR (1000)	STTR (5000)	λ	*A*
r	− 0.15	0.11	0.39	0.45	0.49	− 0.43	− 0.19
p	0.04	0.13	6×10^{-8}	2×10^{-10}	1×10^{-12}	2×10^{-9}	0.01

Lexical diversity turns out to be positively correlated with the poets' birth year, though not for STTR(N) with small values of N[2] (Table 3).

Table 3 shows that the poets' attitude to lexical repetition remains unchanged over time, but topical heterogeneity and the overall size of vocabulary tend to grow. Figure 1 illustrates this for STTR(5000): being a year younger makes a poet's STTR(5000) 2.5 words higher on average.

[1] Note that a negative correlation between STTR on the one hand and λ and *a* on the other hand is expected because larger STTR means more diversity and the opposite is true of λ and *a*.

[2] An increase in lexical diversity is indicated by a **positive** correlation of STTR with birth year and a **negative** correlation of λ and *a* with birth year.

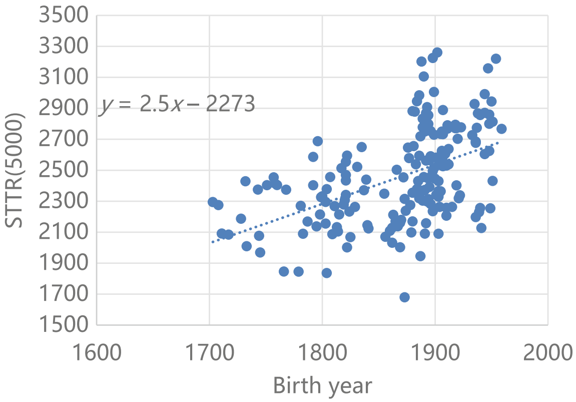


Fig. 1 The relationship between STTR(5000) and birth year

One more sociolinguistic variable that is available from the Russian National Corpus is gender. Since the number of female poets in the eighteenth and nineteenth century is scarce, I limit myself to comparing male and female poets who were primarily active in the twentieth century, which leaves a subsample of 107 male and 16 female poets. It turns out that most lexical diversity measures do not exhibit any significant differences between the two groups, except for λ (Table 4).

An average λ of 0.00079 means an average number of effective words being 1270 for male poets, and for female poets the corresponding number is 1042. Several other studies have reported higher values of lexical diversity in texts by male authors, so this finding is in line with what could be expected [15, 16]. However, it is noteworthy to highlight which measures exhibit a difference between male and female poets. Given that the median value of λ is significantly different for male and female poets and that the same is nearly true for STTR(5000), where the *p*-value is just above the conventional significance threshold of 0.05, one can conclude that these differences are due to the fact that female poets exhibit higher topical heterogeneity than male poets.

However, there is more to this observation. A calculation of λ for male and female poets for different parts of speech independently demonstrates that male poets are significantly more diverse in their use of nouns and verbs, but not adjectives (Table 5).

Table 4 Lexical diversity of male and female poets of the twentieth century

	n	STTR (25)	STTR (100)	STTR (500)	STTR (1000)	STTR (5000)	Λ	*a*
Male	107	19.3	90.8	400.7	729.4	2575.8	0.00079	0.81
Female	16	19.2	90.2	392	706	2417.3	0.00096	0.83
Mann–Whitney *U*		725	725	665	637	594	506	680
P		0.33	0.33	0.15	0.10	0.051	0.009	0.19

Table 5 Lexical diversity (λ) for different parts of speech in male and female poets of the twentieth century

	n	Nouns	Adjectives	Verbs
Male	107	0.00176	0.00315	0.00239
Female	16	0.00223	0.00340	0.00297
Mann–Whitney *U*		480	701	471
P		0.0048	0.33	0.0039

Thus, gender differences in lexical diversity are intertwined with differences in the use of different parts of speech.

4 Conclusion

Lexical diversity of Russian poets can be measured using different metrics, highlighting different aspects of diversity, such as lexical repetition, topical homo-/heterogeneity, and complexity of language and style. The study, based on the texts of 183 Russian poets from the eighteenth century to the present day shows that lexical diversity has been gradually increasing over time, especially with regard to topical homo-/heterogeneity and stylistic complexity. Male authors exhibit higher lexical diversity than female authors with respect to topical homo-/heterogeneity, but this is only limited to nouns and verbs rather than adjectives.

This paper only identifies patterns observed in the data; it does not provide an interpretation or an explanation of why these patterns have arisen, which remains a promising topic for future research.

Acknowledgements The research was supported by the Russian Science Foundation under grant no. 19-78-10081.

References

1. Wachal, R. S., & Spreen, O. (1973). Some measures of lexical diversity in aphasic and normal language performance. *Language and Speech, 16*, 169–181.
2. Malvern, D., Brians, R., Chipere, N., & Durán, P. (2004). *Lexical diversity and language development: Quantification and assessment*. Palgrave Macmillan.
3. Kyle, K., Crossley, S. A., & Jarvis, S. (2021). Assessing the validity of lexical diversity indices using direct judgements. *Language Assessment Quarterly, 18*, 154–170. https://doi.org/10.1080/15434303.2020.1844205
4. Garrard, P. (2004). The effects of very early Alzheimer's disease on the characteristics of writing by a renowned author. *Brain, 128*, 250–260. https://doi.org/10.1093/brain/awh341

5. Le, X., Lancashire, I., Hirst, G., & Jokel, R. (2011). Longitudinal detection of dementia through lexical and syntactic changes in writing: A case study of three British novelists. *Lit. Linguist. Comput., 26*, 435–461. https://doi.org/10.1093/llc/fqr013
6. Jarvis, S. (2013). Chapter 1. Defining and measuring lexical diversity. In: S. Jarvis & M. Daller (eds.) *Studies in Bilingualism* (pp. 13–44). John Benjamins Publishing Company. https://doi.org/10.1075/sibil.47.03ch1
7. Pache, C.O., Dué, C., Lupack, S., & Lamberton, R. (eds.).(2020). Formula. In: *The Cambridge guide to Homer* (pp. 123–125). Cambridge University Press. https://doi.org/10.1017/9781139225649
8. Tognini-Bonelli, E. (2001). *Corpus linguistics at work*. John Benjamins.
9. Burtsev, M., Seliverstov, A., Airapetyan, R., Arkhipov, M., Baymurzina, D., Bushkov, N., Gureenkova, O., Khakhulin, T., Kuratov, Y., Kuznetsov, D., Litinsky, A., Logacheva, V., Lymar, A., Malykh, V., Petrov, M., Polulyakh, V., Pugachev, L., Sorokin, A., Vikhreva, M., … Zaynutdinov, M. (2018). DeepPavlov: Open-source library for dialogue systems. In: Proceedings of ACL 2018, System Demonstrations. pp. 122–127. Association for Computational Linguistics. https://doi.org/10.18653/v1/P18-4021
10. Starchenko, A., & Lyashevskaya, O. (2019). A cross-genre morphological tagging and lemmatization of the russian poetry: distinctive test sets and evaluation. In D. A. Alexandrov, A. V. Boukhanovsky, A. V. Chugunov, Y. Kabanov, O. Koltsova, I. Musabirov (eds.) *Digital transformation and global society*. pp. 732–743. Springer International Publishing. https://doi.org/10.1007/978-3-030-37858-5_62
11. Tweedie, F. J., & Baayen, R. H. (1998). How variable may a constant be? Measures of lexical richness in perspective. *Computers and the Humanities, 32*, 323–352. https://doi.org/10.1023/A:1001749303137
12. Zenker, F., & Kyle, K. (2021). Investigating minimum text lengths for lexical diversity indices. *Assessing Writing, 47*, 100505. https://doi.org/10.1016/j.asw.2020.100505
13. Scott, M. (2004). WordSmith tools.
14. Jarvis, S. (2013). Capturing the diversity in lexical diversity. *Language Learning, 63*, 87–106. https://doi.org/10.1111/j.1467-9922.2012.00739.x
15. Argamon, S., Koppel, M., Fine, J., & Shimoni, A.R. (2003). Gender, genre, and writing style in formal written texts. *Text—Interdiscip. Journal of Study Discourse. 23*. https://doi.org/10.1515/text.2003.014
16. Litvinova, T., Seredin, P., Litvinova, O., & Zagorovskaya, O. (2017). Differences in type-token ratio and part-of-speech frequencies in male and female Russian written texts. In: Proceedings of the Workshop on Stylistic Variation. pp. 69–73. Copenhagen, Denmark: Association for Computational Linguistics. https://doi.org/10.18653/v1/W17-4909

A Semantic Corpus of Russian Literature of 18 Century: Its Current State and Its Future

Marina Ponomareva

Abstract The article describes general principles underlying the document-oriented database "Russian literature of the eighteenth century". The key difference between two approaches to making philological corpora (shallow text vs deep markup) is delineated. Some existing competing projects are briefly summarized and the importance of semantic markup and standard encoding (TEI) is stressed. The article outlines the general principles of multi-facet text markup, lists principal technical capabilities of the database and runs through some text encoding challenges the project is facing. Special attention is drawn to the problem of finding a base form of a series of designations of the same realia (so called reference reconciliation). Cases are identified where such a base form cannot exist and where the reverse process of reference splitting need to be employed. A challenging objective of choosing a proper framework for topical classification of texts is stated. The applicability of the framework to a broader set of texts is asserted.

Keywords Digital humanities · TEI · Russian culture of 18th Century · Document-oriented databases

1 Introduction

One of the most important issues of modern literary studies is that of finding new principles how to represent the history of literature. Despite significant advances in computer technologies, their use in philological studies and especially in literary studies, remains relatively limited. Meanwhile, some tasks of the modern literary studies may only be successfully solved by applying information technologies [1, 17].

M. Ponomareva (✉)
St. Petersburg University, Universitetskaya Emb. 7/9, 199034 St. Petersburg, Russia
e-mail: marina.ponomareva@spbu.ru

P. Eismont et al. (eds.), *Literature, Language and Computing*,
https://doi.org/10.1007/978-981-99-3604-5_11

One example of calling upon modern information technologies to study the literature is the document-oriented database (DOD) “Russian literature of the 18 century”,[1] which is the subject of the present article. It attempts to provide digital versions of essential texts of the 18th-century Russian culture, accompanied by various commentaries, bibliography, learning modules, audiovisual illustrations. However, the primary task of the project is to develop certain new frameworks of philological analysis. Accomplishing such task would allow for a clearer understanding of a series of yet unsolved problems in literary studies and pave the way to creating a radically new view of the literary process, meeting the needs of the modern literary theory.

2 Existing Projects

At present, most digital text collections are just digital copies of their printed originals. Several digital collections targeting specifically Russian literature are well known, among which are **“**The Fundamental Digital Library of Russian Literature and Folklore” (http://feb-web.ru) [26], “Russian Virtual Library” (http://www.rvb.ru) [13, 17], “ImWerden” (http://imwerden.de) [20]. The search capabilities of such libraries are basically confined to hierarchical navigation and standard word-form search.[2]

Existing digital collections mostly tend to reproduce traditional forms of textual organization, the text being inextricably associated with its comprising edition. So the possibilities of modern technologies remain unclaimed, since the only new method of searching (compared to printed books) is a word-form search. All other relations that may exist between texts (based on their date of publication, metrical principles, topics, dedications) are for the most part ignored, as neither printed books, nor their simple digital copies explicate them. So the main goal of such digital libraries is to provide access to the text of a book (by itself being of course the most noble task), and not to organize scholarly work involving it. One may argue that it’s the latter task that would become a mainstream in the further development of digital libraries, and the philology as a whole would benefit from the progress thereof, since the new ways to represent texts would hopefully lead to new insights onto their essential characteristics [8].

[1] The project of the DOD “Russian literature of the eighteenth century” (https://18vek.spb.ru/) is one of the projects of an interdisciplinary research seminar “Russian eighteenth century” which brings together scholars from Saint-Petersburg State University, Institute for linguistic studies (RAS), Institute of Russian literature (RAS). The DOD has been created by teachers and students of SPbSU since 2007. In 2011–2014 the project was funded by RFBR.

[2] For a detailed review of existing digital collections, see [3].

There are, however, projects with more advanced capabilities, naming only a few[3]:

- Bocaccio's Decameron: http://www.brown.edu/Departments/Italian_Studies/dweb/index.php
- Emblem Project Utrecht: https://emblems.hum.uu.nl/
- Bibliothéques Virtuelles Humanistes: http://www.bvh.univ-tours.fr/ [10]
- the annotated corpus of Russian literature for children DetCorpus: http://detcorpus.ru [14]
- the corpus of Russian and Ukrainian diaries: http://prozhito.org [12, 16]
- Avtograf: a digital archive of twentieth century Russian literature (http://literature-archive.ru/)
- The project of a digital critical edition of works by A. Pushkin: https://www.pushkin-digital.ru/
- The project of a "semantic" edition of works by Leo Tolstoy (http://tolstoy.ru/projects/tolstoy-digital/) [5][4]

The texts in such resources are marked up in a special way, which allows not only searching by word-forms, but also by author, by publication year, by topic, etc. This is the kind of corpora that we refer to as "semantic" because they enrich the actual source texts with the results of scholarly analysis.[5] However, in most of the corpora listed above the mark-up is done on the level of whole texts, so e.g., it is impossible to find exactly those pieces of a text that mention a given person. Also it should be noted that, while many such projects in Europe predate ours by a decade or so, Russian projects of that kind appeared much later than our project.[6]

3 Principles of Multi-facet Text Encoding

Creating our DOD, we intended to represent the 18th-century Russian literature as a consolidated textual field of a sort, where each work is potentially linked to a whole series of other works.

Using the system as a tool to describe the history of literature, we employ the possibilities that are provided by modern computer technologies to work with literary texts. First of all, we mean multidimensional analysis by a set of parameters. Each text in the system is characterized by a number of established parameters. These include both formal textual characteristics, and substantial ones, that is, those that do require

[3] To the best of our knowledge, there is no exhaustive list of such projects. The TEI consortium maintains a list of projects that use TEI in some way (see https://tei-c.org/activities/projects/), however, the list is neither complete, nor up to date, and besides, it includes all kinds of projects, not only philological ones.

[4] This project seems to be closest in spirit to what we're doing, but unfortunately it seems to be not available to general public as of now.

[5] Another common term for such corpora is "deep markup," see, for example, [15].

[6] The detailed comparison of various semantic corpora would certainly deserve a separate article, so the provided list should be considered purely illustrative.

interpretation. Among the most important parameters shall be listed the author, the publication and creation dates, the publication location, the rhyming scheme and the poetic meter for verses, the genre, a set of topics and a set of proper names being mentioned. So a user can select a group of texts for further study, matching the selected criteria.

Due to such multi-facet text description, each work gets linked to others along several axes. And this lays the ground for a new representation of the literary process: transforming the corpus according to a view, formed by a set of parameters, a scholar may regard the literary life of the eighteenth century from several viewpoints, either chronologically, or by groups of topics, or by groups of people or places being mentioned and so on.

The markup of our corpus complies with TEI guidelines (*Text Encoding Initiative*, http://www.tei-c.org) [23–25]. P4 version is currently used, but plans do exist to migrate the corpus to a newer and more flexible P5. Initially we were using the SGML variant of TEI, not XML because classic SGML is more compact and thus more suitable for manual marking-up [4]. However, over the past decade, SGML has finally got out of fashion, and at the same time, an abundance of XML editing tools has emerged, so now we have switched to XML as well [19]. We strive to follow TEI guidelines strictly, not adding new elements or attributes, however we impose additional constraints on the allowed set of values for certain attributes where TEI itself leaves that under-specified [7]. Also it should be noted that, due to the vast diversity of the information being encoded, we use TEI at its full capacity, which may later somewhat impede interaction with other projects that are limited to a more basic subset of TEI.

The software behind the DOD is itself free/libre and is based solely on free/libre components (the repository of our project at GitHub: https://github.com/antology-xviii). The core of the system is written in OCaml and uses any suitable RDBMS to store indexing data (though historically other less common formats were also tried, such as HDF [21] or RDF triple store [2, 18]). The general process of conversion between XML sources and the indexing database follows the ideas outlined in [22]. Our OS of choice is currently Debian GNU/Linux 11, however most (if not all) components of the system are fully platform-neutral.

Thanks to the open (in all senses) software architecture, the possible set of search parameters is not rigid, parameters may be added or amended as needed. This is essential for philological studies, as it is indeed for any scholarly work. The process of exploring the object (a text in our case) reveals some previously unknown (or even considered impossible) features and draws attention to some other features that were regarded as irrelevant. Thus, we regard the possibility of *incorporating* such newly-discovered features back into the system as one of its key points.

A text, once described, does not stay immutable forever: it may be enriched with new characteristics, should they turn out to be necessary for further work. Since the encoding of texts in our system is open, individual scholars may tune it to their private needs, to solve their own tasks.

Such a framework opens up new horizons of working with texts. Before the scholar had to re-establish various relations between literary works each time afresh in their head, now these relations may be fully explicated and made available for everybody to study.

4 Normalization Challenges

Now we will outline some challenges that we have faced and the ways to solve them.

The evolution of our system was shaped primarily by the evidence, that is, by the texts under study. So in the beginning there was an approximate list of parameters that were considered necessary to describe our corpus properly. However, soon the list started to get amended and extended. For example, the initial features "toponyms" and "personal names" were subcategorized into groups: names of real places and historical figures, mythological names, Biblical names and names of fiction.

In the same line, the ground principles of feature assignment were constantly being refined, and it concerns not only substantial parameters but also such formal ones as/ for example, those dealing with metrics.

One of the most important recurring issues that we faced while preparing the texts for the DOD is the issue of choosing the base form among a set of synonymous names of the same phenomenon. This can be called the problem of lemmatization, taking this term in a broad sense to cover synonym unification and plain linguistic word-form unification.[7]. Let's regard the case of searching for base forms for geographic names.

For example, the Black Sea may be called *Pont*, *Evksin*, *Evksinskiy Pont,* or *Chyornoye more* in the texts of the eighteenth century, so the set of names correspond to a single physical entity. To ensure the search completeness, a single canonical name should be chosen to identify the entity in the DOD.

Our initial approach was to use a modern conventional name for such cases.

Thus, the base form for the following:

- *Pont*, *Evksin*, *Evksinskiy Pont, Chyornoye more* → *Chyornoye more*
- *Belt* → *Baltiyskoye more*
- *Rifey*, *Rifeyskiye gory* → *Ural, g.*

However, the case of a series *Rossiya—Rus'—Rossiyskaya imperiya — Rossiyskoye gosudarstvo*, including also a metonymical designator *Sever* (the North) is more complicated, as it is far from clear whether indeed all these names have the same extension, and to what degree, so whether it's not a failure to ascribe them to the same base form.

[7] A more exact term adopted in the field of information retrieval theory may be *reference reconciliation*, see [11]

Such an approach—using a modern name as the base form—does have its own merits and deficiencies. In particular, any user, not necessary an expert in the eighteenth century culture, can find realia she is interested in, discover naming variability and see the usage contexts, by starting with a common modern name. However, should the scholars be only interested in contexts featuring just some specific name form (i.e. *Evksin*), they would be urged to perform a selection of contexts on their own behalf.

Besides the approach somewhat fails when the entity in question just does not have a properly modern name because it does not exist anymore, for example cf. such designations as *Kirgiz-Kaysatskaya orda.*

Musing over the issue, we concluded that the initial decision to use the modern name as the base form needs to be revised. The more correct method turns out to be the principle of *double* designation, which includes both the modern name (*where it is applicable*) and the real form, as seen in the text. In this way, the search becomes more flexible and complete, and the problem of finding a base form for the eighteenth century specific phenomena disappears completely.

The cases of un-lemmatization deserve special mentioning, when several forms are explicitly *not* consolidated, or even when the same surface name is logically split into several types.

The waiver of the base form is employed when describing characters of antique mythology—the variant (a Greek or a Roman one) is always preserved that was used by the author.

To complicate the matter further, some names are ambiguous by their nature. As a rule, those are the names of loci, holding a double reference—a real geographical object and its mythological or fictional representation:

For example:

1. *Olimp* (Olympus) as (a) the name of a real mountain; (b) the name of a place in mythology
2. *Arkadia* as (a) the name of a real geographical region; (b) the fictional name
3. *Petropol'* was initially interpreted as a synonym for Saint-Petersburg, however later it was decided to keep it separately as a purely literary name of the city, full of antique connotations.

In cases like (1) and (2) the base form is constructed from the textual form augmented with its classifier, e. g. *Olimp (myth.)* versus *Olimp (real).*

Any kind of scholarly work, and especially the analysis of works of fiction, forces scholars to change their initial picture of the object being studied according to its inherent properties. Therefore our starting selection of base forms for the DOD turns out to be far from adequate, and still we have no full confidence that e.g. our current solution for geographical names is the final one.

Even more complex is the task of *topical description* of texts, in particular, the choice of a base form for the set of designations of the same topic. For us this is work in progress; one may argue that a proper solution to this problem requires turning to more formal logical frameworks [9].

We may hope that, though our framework is based on a relatively small corpus, it should be possible to generalize it later to cover other periods of Russian literature as well as the set of works by a single author.

References

1. Andreev, A. V., Bukharkin, P. E., Matveev E. M., & Ponomareva, M. V. (2009) O razrabotke novoi teoreticheskoi modeli reprezentatsii istorii literatury (na materiale russkoi literatury XVIII veka) [On developping a new theoretical model of representation of literary history] (in Russian). *Literaturnaia kul'tura Rossii XVIII veka [Russian literary culture in 18th century]* (3), 303–310
2. Andreev, A. V. (2012). Ispol'zovanie TEI i RDF v informatsionno-poiskovoi sisteme "Russkaia literatura XVIII veka" [Using TEI and RDF in the informational system "Russian literature of 18th century"] (in Russian). In V. A. Baranov, A. G. Varfolomeev (eds.) *Informatsionnye tekhnologii i pis'mennoe nasledie: Materialy IV mezhdunarodnoi nauchnoi konferentsii El'Manuscript–2012* [Digital Heritage. The proceedings of El'Manuscript–2012], Petrozavodsk, 03–08 sentiabria 2012 goda, pp. 11–12. Petrozavodsk
3. Andreev, A. V., & Ponomareva, M. V. (2014). Sravnitel'nyi analiz podkhodov k sozdaniiu elektronnykh filologicheskikh kollektsii [A comparative study of methods to create digital philological collections] (in Russian). *Literaturnaia kul'tura Rossii XVIII veka [Russian literary culture in 18th century]* (5), 295–305
4. Barnard, D. T., Burnard L., & Sperberg-McQueen, C. M. (1996). Lessons learned from using SGML in the Text Encoding Initiative. *Computer Standards & Interfaces 18* (1), 3–10. https://doi.org/10.1016/0920-5489(95)00035-6
5. Bonch-Osmolovskaia, A., Kolbasov, M., Orekhov, B., Pavlova, I., & Skorinkin, D. (2018) Semanticheskoe izdanie tekstov L. N. Tolstogo: ot teksta k ontologii [The semantic edition of Leo Tolstoy's texts: from text to ontology] (in Russian). *Napis* (XXIV), 381–391. https://rcin.org.pl/dlibra/show-content/publication/97547?id=97547. Last Accessed 14 May 2022
6. Burnard, L (1999) Is humanities computing an academic discipline? or, why humanities computing matters. In *Presented at an interdisciplinary seminar at the Institute for Advanced Technology in the Humanities*. University of Virginia. http://www.iath.virginia.edu/hcs/burnard.html. Last Accessed 14 May 2022
7. Burnard, L. (2019).What is TEI conformance, and Why should you care?. *Journal of the Text Encoding Initiative* (12). https://doi.org/10.4000/jtei.1777
8. Buzzetti, D. (2002). Digital representation and the text model. *New Literary History 33*(1), 61–88
9. Ciotti, F., & Tomasi, F. (2016). Formal ontologies, linked data, and TEI semantics. *Journal of the Text Encoding Initiative* (9). https://doi.org/10.4000/jtei.1480. https://journals.openedition.org/jtei/1480
10. Demonet, M. L. (2009). The Bibliothèques Virtuelles Humanistes (Virtual Humanistic Libraries) in Tours: a Collection, or a Corpus? In *Digital Humanities Meeting*. University of Maryland. http://www.bvh.univ-tours.fr/presentation_en.asp. Last Accessed 14 May 2022
11. Dong, X., Halevy, A., & Madhavan, J. (2005). Reference reconciliation in complex information spaces. In *Proceedings of the 2005 ACM SIGMOD international conference on Management of data (SIGMOD '05)* (pp. 85–96). Association for Computing Machinery. https://doi.org/10.1145/1066157.1066168
12. Drapkin, I., & Tyshkevich, N. (2017). Prozhito: Private diaries database. In *DHN 2017 Digital humaniora i Norden/Digital Humanities in the Nordic Countries* (pp. 164–165). http://dhn2017.eu/wp-content/uploads/2017/03/DHN2017_Book_of_Abstracts_20170313.pdf. Last Accesed 14 May 2022

13. Litvinov, V., & Pilshchikov, I (2004). Russian Virtual Library (RVL): Some aspects of visualization of literature text and comments to it. In EVA 2004 Berlin: Konferenzband: Elektronische Bildverarbeitung & Kunst, Kultur, Historie: 10.-12. November 2004 in den Staatliche Museen zu Berlin am Kulturforum Potsdamer Platz: Die 11. Berliner Veranstaltung der internationalen EVA-Serie Electronic Imaging & the Visual Arts, pp. 124–126
14. Maslinskii, K., Lekarevich, E. & Aleinik, L. (2021) Korpus russkoi prozy dlia detei i iunoshestva. [The corpus of Russian prose for children and young adults] (in Russian) In *Repozitorii otkrytykh dannykh po russkoi literature i fol'kloru* [The repository of open data of Russian literature and folklore] (V1) (2021). https://doi.org/10.31860/openlit-2021.4-C001
15. McCarty, W. (2003). Depth, markup and modelling. *Text Technology 12*(1), 59–74
16. Mel'nichenko, M. A. & Tyshkevich, M. A. (2020). Prozhito ot rukopisi do korpusa [Prozhito: from manuscript to corpus] (in Russian). *Elektronnyi nauchno-obrazovatel'nyi zhurnal "Istoriia"* [Digital educational magazine "History"] *7*(61), 2. https://history.jes.su/s207987840001935-7-1. Last Accessed 14 May 2022
17. Mjør, K. J. (2009). The online library and the classic literary canon in Postsoviet Russia: Some observations on "The Fundamental Electronic Library of Russian Literature and Folklore". *Digital Icons* (1, 2), 83–99
18. Morbidoni, C., Pierazzo E., & Tummarello, G. (2005) Toward textual encoding based on RDF. In *Proceedings of the 9th ICCC International Conference on Electronic Publishing* (pp. 57–64). Leuven-Heverlee
19. Nellhaus, T. (2001). XML, TEI, Digital libraries in the humanities. *Portal: Libraries and the Academy 1*(3), 267–277
20. Nikitin-Perenskii, A. (2009).O novykh postupleniiakh v elektronnuiu biblioteku «ImWerden» [Some new acquisitions of the digital library ImWerden] (in Russian). *Toronto Slavic Quartetly* (29). http://sites.utoronto.ca/tsq/29/nikitin29.shtml. Last Accessed 14 May 2022
21. Pellegrino, D. A. (2009). *Making novel connections between literature and data.* http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.457.9410&rep=rep1&type=pdf. Last Accessed 14 May 2022
22. Simons, G. F. (1999). Using architectural forms to map TEI Data into an object-oriented database. *Computers and the Humanities, 33*, 85–101. https://doi.org/10.1023/A:1001765030032
23. Skorinkin, D. A. (2016) Elektronnoe predstavlenie teksta s pomoshch'iu standarta razmetki TEI [Digital text encoding using TEI Guidelines] (in Russian). Vestnik Moskovskogo universiteta. Seriia 9: Filologiia [Moscow University Bulletin. Series 9. Philology] (5), 90–108
24. Sperberg-McQueen, C. M. (1994). The text encoding initiative: electronic text markup for research. In B. Sutton (ed.) *Literary Texts in an Electronic Age* (pp. 35–55). Urbana-Champaign, IL: University of Illinois at Urbana-Champaign, Graduate School of Library and Information Science
25. Vanhoutte, E. (2004). An introduction to the TEI and the TEI consortium. *Literary & Linguistic Computing 19*(1), 9. https://doi.org/10.1093/llc/19.1.9
26. Vigurskii, K. V., & Pil'shchikov, I. A. (2005). Fundamental'naia elektronnaia biblioteka «Russkaia literatura i fol'klor» (FEB) [**T**he Fundamental Digital Library of Russian Literature and Folklore] (in Russian). *Voprosy internet-obrazovaniia* [Issues of Intenet education] (27). http://vio.uchim.info/Vio_27/cd_site/articles/art_1_16.htm. Last Accessed 14 May 2022

Computing in…… Language Teaching and Translation

Multimedia Dictionary of Verbal Vocabulary: Concept, Structure, Implementation

Elena R. Laskareva and Alina A. Pozdnyakova

Abstract Being one of the most complex parts of speech, the verb requires a comprehensive approach in the lexicographic description in different subtypes of dictionaries. An appropriate verb dictionary that offers various ways of semantization helps better understand of the meaning of a verb by a student. This article discusses the requirements for an electronic verbal dictionary as a component of the Russian language teaching system for foreign students. At the moment, there is no resource in which all verb minimums for different stages of mastering the Russian language could be brought into a single system. This gap can be filled by an electronic verbal dictionary with searching functionality allowing (1) to show verbs in structural, semantic and functional aspects; (2) to present verbal units as constituent elements of the language system; (3) to utilize both the information and training potential of electronic educational resources on one platform. In solving these problems, the article discusses the general concept of such a dictionary and particular issues of its implementation (the structure of the dictionary, the content and structure of the dictionary entry, a set of relationships between verbs, etc.).

Keywords Electronic lexicography · Verbal dictionary · Dictionary entry · Russian as a foreign language

1 Introduction

The reference to the verb in our work is not accidental. The verb acts as the organizing center of a sentence in any language. In combination of the verb with other words, the specificity of the language, its national flavor, is clearly manifested. Without deep study of the verbal system, it is impossible to understand a foreign language,

E. R. Laskareva (✉)
St. Petersburg University, Universitetskaya Nab. 7/9, St. Petersburg 199034, Russia
e-mail: e.laskareva@spbu.ru

A. A. Pozdnyakova
The Kosygin State University of Russia, 33/1, Sadovnicheskaya Str, Moscow 117997, Russia

P. Eismont et al. (eds.), *Literature, Language and Computing*,
https://doi.org/10.1007/978-981-99-3604-5_12

so the task of many lexicographic studies has always been and remains, to identify derivational and syntagmatic connections of verbs of various lexical and semantic groups. In this regard, the search for a certain tool for consolidating “resources of knowledge” and the introduction of digital educational technologies for these purposes is of particular importance today.

Currently, universities actively apply digital technologies in their educational processes. Their application in the process of teaching Russian as a foreign language allows achieving multiple important tasks related to the organization, didactic and methodological support of both classroom and home work of students. The use of electronic textbooks and manuals, electronic dictionaries and databases, teaching and controlling computer programs, various applications as a supplement to traditional study material, allows students to intensify the process of learning a foreign language. It gives them the opportunity to choose an individual strategy for mastering the study material, monitors their own achievements at various stages of work and identifies skills that need correction. According to Pastor and Alcina [1], “the use of electronic dictionaries has many advantages over the traditional paper dictionary” [1: 1]. However, access to the lexicon and terminology of a dictionary (and other electronic resources) “presents certain difficulties, partly due to the lack of user knowledge (even among language experts such as translators) about how a dictionary can be queried to access such information, and partly due to the diversity of ways a dictionary can be consulted (in different areas of the dictionary, with different operators, in widely varied interfaces) which vary from one dictionary to another” [1: 1].

The problem of creating electronic educational dictionaries is relevant, but inadequately developed in scientific terms. Educational lexicography in its electronic form in most cases is limited to a narrow range of purely applied tasks (creating glossaries, verb minima for courses, dictionary comments to assignments, etc.) and is desultory. Available electronic resources have other negative aspects:

(1) the methodological grounding of educational dictionaries is not always based on the data of psychology and physiology, so the mechanisms of perception and assimilation of digital material by specific categories of users are ignored;
(2) the content of dictionaries is maximally standardized irrelevantly to a specific ethnic scientific school of the user;
(3) the structure of dictionaries is incomprehensible for foreigners studying Russian, the language of their specialty.

The purpose of this work, as we see it, is to present the concept of a multimedia verb dictionary that considers the needs of a foreign student studying Russian as much as possible.

2 Methods

While finding a solution to the problem of creating an “optimal verbal dictionary,” we have used the data of modern computer lexicography [1–3] and use NLP (Natural Language Processing) methods [4–6]. Based on this position, we single out such basic characteristics of the electronic verbal dictionary as the multiplatform orientation of the operating system (OS platforms Linux, Solaris, Windows), multilingualism, multimedia, and the interactive content.

We have analyzed the market of existing electronic dictionaries to identify trends and perspectives for its development, considering the subject-professional content orientation. Electronic versions of printed lexicographic publications, on the one hand, and dictionaries created specifically for use in electronic format, on the other hand, form a common information technology lexicographic field. The specificity of the second group is the availability of wide additional functionality provided by the computer code on the basis of which dictionaries are created. According to Sven Tarp [7], “the traditional bilingual dictionary which contrasts two different languages is inconvenient. This kind of dictionary usually provides too little data to fully assist the users in foreign-language text production. And if it does furnish the needed data, the inevitable result would be that many articles would be filled with too much data, thus creating a new problem for its users, namely information stress due to data overload” [7: 409]. Therefore, obviously, we should talk about a lexicographic resource of a new type. Vast majority of researchers tend to think that an electronic computer dictionary needs a computer algorithm that allows users of various backgrounds (from ordinary native speakers to scientists) to exploit the functionalities of the vocabulary base of each particular vocabulary genre in the most efficient manner.

To date, two alternative approaches to the analysis of material have been defined in computer lexicography—formal (logical) and linguistic. The first (formal) is based on logic (first-order predicates, descriptive, modal, etc.) [8, 9]. The second (linguistic) is based on the study of natural language (lexicon, grammar), analysis of the language corpus, and construction of thesauri based on it [4, 5, 10]. Combination and “intersection” of approaches allow to create resources with elements of formal axiomatics and logical systems with inclusions of linguistic knowledge [11, 12]. The verb dictionary we offer also belongs to the latter type.

3 Results

The analysis of linguistic, psychological and methodological literature leads us to the conclusion that the process of assimilation of verbal vocabulary by a non-native speaker has a number of specific characteristics that must be considered when working on a dictionary. These characteristics (linguistic, cognitive, motivational) form certain “contexts” that help or hinder the acceptance of an “alien” verbal system, its “cognitive structuring and activation in speech.

Difficulties that arise in the development of verbal vocabulary by foreign students are primarily associated with differences in the aspect meanings in the contacting languages and, as a result, with a lack of understanding of the peculiarities of the use of Russian perfective and imperfective verbs. A separate problem is the verbal prefixation, which is closely related to the expression of the aspect category in Russian and therefore needs a special lexicographic description. Being a traditional means of expressing perfect forms, the Russian verbal prefix does not always “fit” into the grammatical systems of other languages (Turkic, Semitic, etc.), which creates difficulties for Russian language learners in understanding the aspect specific meanings of verb constructions and, accordingly, their adaptation in translation. Only knowing and competently using the general and particular patterns of the process of assimilation of verbal material, the compiler of the dictionary can achieve effective results of his activity.

The presented verbal dictionary summarizes the experience of compiling educational dictionaries accumulated by Russian and foreign science [13–16]. As for its orientation, it is an educational dictionary. From the viewpoint of the sphere of inclusion, this is a lexico-grammatical dictionary. The dictionary includes the most common verbal units and their derivatives that are relevant for a foreigner at each specific stage of learning the Russian language.

When choosing the verbal minimum, we started from the fact that the selected vocabulary should reflect the central structural and semantic features of a particular verb group, on the one hand, and be accessible to foreign students, on the other. When selecting verb units, we applied the generally accepted principles of vocabulary selection, such as linguistic value, educational and methodological expediency, and frequency of use.

Working on the dictionary project, we used the following algorithm for presenting material:

1. A motivating non-prefix verb is given as a head verb. Its semantization can be carried out by the verbal unit of the user’s native language (when the appropriate option exists).

2. Prefixed verbal derivatives are given in alphabetical order, which makes it easier to find them when determining their meaning. The semantization of derivatives is carried out in accordance with the user's native language (when the appropriate option exists).
3. Groups of prefixed verbal derivatives are given in pairs of imperfective and perfective form, which allows you to immediately present the complex variants of verb forms. Since in many foreign languages both aspects are presented in one verbal unit, it is possible to determine the meaning of the aspect in context. Properly chosen examples make it possible to fully show the aspect and tense meaning that a given prefixed verb can express.
4. Verbs are given as part of syntactic constructions, which can be presented as schemes of phrases or sentences and contain an indication of the types of sentences in which prefixed verbs are most frequently used. Here, comments in the user's native language are possible, backing similar Russian constructions.

To display the linguistic structure in the dictionary, the nesting principle is used, which allows to most fully reflect the systemic connections of single-root words (verbs, verbal nouns, adjectives, etc.) [17].

The dictionary is on a block construction principle. The lexical base of each block is made up of verbal units of a specific lexical and semantic/thematic group. As motivators, we single out non-prefixed verbs of several lexico-semantic groups, which give the largest number of prefixed derivatives. Among them, the following seemed particularly relevant to us: (1) verbs of speech; (2) verbs of thought, intellectual activity; (3) verbs of perception; (4) verbs of motion; (5) verbs of position change in space; (6) verbs of labor activity; (7) verbs of feeling and state; 8) verbs of behavior; (9) verbs of being and existence [18].

The disclosure of the meanings of verbs and their derivatives in a dictionary entry is carried out in stages, through a system of hyperlinks of different levels, for each lexical unit separately.

A hyperlink in our work is an option that allows you to go to a specific page of the dictionary: to the morpheme page, showing its meanings and relationships, or to the page of a derivative word, demonstrating its meaning in context.

Visually, the system of such links can be depicted as a tree, in which the main page containing the motivating word (descriptor) is the trunk, and the branches are nested documents of level 1, 2, 3, 4, etc. The nesting levels are shown in Fig. 1.

When determining the word-formation links of the motivating word, the data of Tikhonov [19] was used, while the interpretation is based on existing verbal dictionaries (including bilingual, educational) and modern research [13, 20, 21].

For example, for units of the group of verbs of being, the system of hyperlinks may look like this:

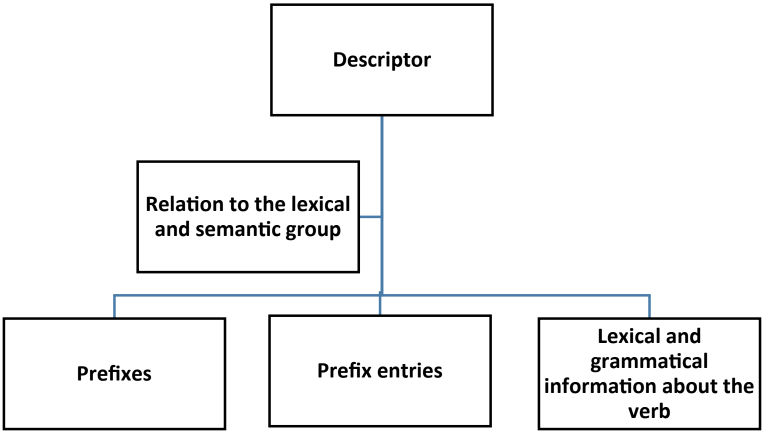

Descriptor
Relation to the lexical and semantic group
Prefixes
Prefix entries
Lexical and grammatical information about the verb

Fig. 1 Nesting levels (1–4) of elements of a dictionary entry

Motivating Word (Descriptor)

Byt'.

Hyperlinks of the 1st Level (Relation to the Lexical and Semantic Group)

Byt' kem? chem? v kom? v chjom?, **zhit'** kem? chem? s kem? na chto? v kom? v chjom?, **sushhestvovat'** chem? na chto? v kom? v chjom?, **imet'sja** u kogo? v kom? v chjom?, **nahodit'sja** u kogo? v kom? v chjom?, **vodit'sja** v kom? v chjom?

Hyperlinks of the 2nd Level (Prefixes)

Vybyt'—vybyvat', dobyt'—dobyvat', zabyt'—zabyvat', zabyt'sja—zabyvat'sja, otbyt'—otbyvat', pobyt', pobyvat', prebyvat', pribyt'—pribyvat', probyt', ubyt'—ubyvat'.

Hyperlinks of the 3rd Level (Prefix Entries)

Dictionary entries devoted to prefixes are built according to a general principle and include: (1) the name of the prefix; (2) description of its semantics based on the traditional approach; (3) a model of verbs formed with the appropriate prefix (G1).

Hyperlinks of the 4 Th Level (Lexical and Grammatical Information About the Verb)

- **výbyt' (sov.): výbudu, výbudesh', výbudut *nesov.* vybyvát' (iz + II) G1 G7 výbyt' iz góroda; výbyt' iz igrý; výbyt' iz strója**
- **vybyvát' (nesov.): vybyváju, vybyváesh', vybyvájut *sov.* výbyt'** G1 G7
- **dobýt' (sov.): dobyúdu, dobyúdesh', dobyúdut *nesov.* dobyvát' *1.* G1 G7 dobýt' núzhnye svédenija; *2.* G7 dobýt' neft'**
- **dobyvát' (nesov.): dobyváju, dobyváesh', dobyvájut *sov.* dobýt'** G1 G7
- **zabýt' (sov.): zabúdu, zabúdesh', zabúdut *nesov.* zabyvát' *1.* (+ IV, o + VI, + inf.) G1 G7 zabýt' o zasedánii; zabýt' o próshlom; sebjá ne zabúd'! *2.* G7 vy opját' zabýli biléty**
- **zabyvát' (nesov.): zabyváju, zabyváesh', zabyvájut *sov.* zabýt'** G1 G7
- **zabýt'sja (sov.): zabúdus', zabúdesh'sja, zabúdutsja *nesov.* zabyvát'sja *1.* G1 G7 zabýt'sja na minútu; *2.* G7 zabýt'sja tjazhélym snom; *3.* G7 vstrécha ne zabúdetsja; *4.* G7 ne zabyvájsja, jéto bestáktno!**
- **zabyvát'sja (nesov.): zabyvájus', zabyváesh'sja, zabyvájutsja *sov.* zabýt'sja** G1 G7
- **otbýt' (sov.): otbúdu, otbúdesh', otbúdut *nesov.* otbyvát' *1.* G1 G7 póezd otbýl v sem' chasóv; *2.* G7 otbýt' nakazánie; otbýt' vóinskuju povínnost'**
- **otbyvát' (nesov.): otbyváju, otbyváesh', otbyvájut *sov.* otbýt'** G1 G7
- **pobýt' (sov.): pobúdu, pobúdesh', pobúdut G1 G7 my pobýli v Lóndone dva dnja**
- **pobyvát' (sov.): pobyváju, pobyváesh', pobyvájut (v + VI) *1.* G1 G7 on pobyvál vsjúdu; v próshlom godú my pobyváli v Norvégii i v Shvécii; *2.* G7 mne nádo pobyvát' v kontóre**
- **prebyvát' (nesov.): prebyváju, prebyváesh', prebyvájut *1.* G1 G7 on prebyvál v Moskvé dva mésjaca; *2.* G7 prebyvát' u vlásti**
- **pribýt' (sov.): pribúdu, pribúdesh', pribúdut *nesov.* pribyvát' *1.* G1 G7 póezd príbyl vóvremja; *2.* G7 pribylá vodá v reké**
- **pribyvát' (nesov.): pribyváju, pribyváesh', pribyvájut *sov.* pribýt'** G1 G7
- **probýt' (sov.): probúdu, probúdesh', probúdut G1 G7 on próbyl u nas tri nedéli**
- **ubýt' (sov.): ubúdu, ubúdesh', ubúdut *nesov.* ubyvát' *1.* G1 G7 vodá v reké ubylá; interés k voprósu úbyl; *2.* G7 ubýt' v komandiróvku; ubýt' v ótpusk**
- **ubyvát' (nesov.): ubyváju, ubyváesh', ubyvájut *sov.* ubýt'** G1 G7

The following are indicated as dictionary marks: (1) the aspect of the verb–imperfective (nesov.) and perfective (sov.); (2) 1SG, 2SG and 3PL forms of the verb; (3) features of verb government (oblique case (Genitive—II, Dative—III, Accusative—IV, Instrumental—V, Prepositional—VI) and prepositions), (4) stressed syllables of polysyllabic words (otbýt'; vodá).

Assuming the educational nature of the dictionary and the language specifics of the audience it is aimed at, it is reasonable to include a limited number of verbal derivatives in the dictionary entry (hyperlinks of the 4th, 5th, 6th, etc. levels) [19].

Hyperlinks of the 5th Level (Verbal Nouns)

- **Byl'** **from** ***byt'***
- **Bytie** **from** ***byt'***
- **Bytnost'** **from** ***byt'***

Hyperlinks of the 6th and 7th Levels (Verbal Nouns)

- **Pobyvka** **from** ***pobyvat'***
- **Vybyvanie** **from** ***vybyvat'***
- **Vybytie** **from** ***vybyt'***
- **Otbyvanie** **from** ***otbyvat'***
- **Otbyvka** **from** ***otbyt'***
- **Pribyl'** **from** ***pribyt'***
- **Pribytie** **from** ***pribyt'***
- **Ubyl'** **from** ***ubyt'***
- **Ubytok** **from** ***ubyt'***

Hyperlinks of the 5th Level (Verbal Adjectives)

- **Byloj** **from** ***byt'***
- **Byvshij** **from** ***byt'***

Hyperlinks of the 6th and 7th Levels (Verbal Adjectives)

- **Byvalyj** **from** ***byvat'***
- **Sbyvchivyj** **from** ***sbyvat'sja***
- **Ubytochnyj** **from** ***ubyt'***

To activate the logical connections between the elements of the Russian verbal system and the system of the user's native language, it is advisable to enter the most frequently used combinations into the dictionary entry: (a) with a verb and a dependent noun (*dobyt' neft'*); (b) with a verbal derivative and a dependent noun (*dobycha nefti).*

Participles are not included in the dictionary as independent entries, they are activated by means of hyperlinks and the "graphic assistant" function (G7).

The multimedia component of the dictionary implies the possibility of automated search, the availability of internal links between entries, the use of different types of information (not only textual, but also audiovisual) [22–26].

With this approach, the electronic verbal dictionary is "personalized" as much as possible, and it becomes a constantly replenished and updated source of information, and therefore practically keeps its relevance [27, 28]. Thus, the slogan Atkins [29] is implemented: "It is up to us to take up the real challenge of the computer age, by asking not how the computer can help us produce old-style dictionaries better, but how it can help us create something new: to look at the needs of dictionary users of every language, and every walk of life, users as diverse as people themselves, and give them the kind of information they need for whatever they are using the dictionary for, and not simply the popular selection of facts that will pack semi-legibly inside book covers" [29: 516].

4 Discussion

In connection with the lexicographic analysis of verbs, the question arises which of these grammatical features should be recorded in educational dictionaries and how much it is possible to do this without violating the principles of "economy" and "simplicity" (though with good intentions—for the sake of observing the principles of "completeness" and "efficiency" of the description of verb units).

The systems of grammatical markings for head verbs, analyzed according to the leading explanatory dictionaries of the modern Russian language and foreign dictionaries, allow us to fully present the design of the verbal meanings of words [28].

In accordance with the lexicographic tradition, the paradigm of the Russian verb includes both personal and impersonal forms. The verb entry indicates the forms of the 1st and 2nd person singular, the 3rd person plural, stressed syllables, alternations, in some cases—the forms of the past tense and the imperative.

Participles as independent entries are not included in the dictionary (this, however, may not apply to participle adjectives and nouns, for example, *byloj, byvshij, byvalyj).* For other languages (for example, Turkish), the participle is a separate part of speech, therefore, in explanatory dictionaries, each participle is given in a separate dictionary entry.

For Russian verbs, the most relevant marks are associated with the category of aspect, voice and transitivity, mood, which already have a semantic component. Additionally, references to participle forms are expedient, which require a special study of the verbal complex of grammatical meanings.

As additional information that goes beyond inflection, the dictionary entry of a Russian verb provides information about the formation of a verb of the opposite aspect. The paradigm of the transitive imperfective verb also includes passive forms ending in -*sia* (personal and impersonal). At the same time, forms in—*sia* with a non-passive meaning are considered to belong to a separate word—a reflexive verb.

The expediency of introducing this or that audio content into the dictionary was confirmed experimentally. In the course of the experiment, carried out in stages in groups of foreign students studying at different faculties and having a confirmed level of Russian language proficiency not lower than B1, we found that the recognition rate of prefixed verb units in sounding speech turned out to be quite low. The range of recognition of prefixed verbs in sounding speech for all analyzed groups varied

Kt = 0.34−0.72, the range of recognition of verbal nouns was Kt = 0.12 −0.28. The recognition coefficient in the printed text turned out to be expectedly higher, but it cannot be considered sufficient for obtaining professional education in Russian: for verbs Kt = 0.44−0.78, for verbal nouns Kt = 0.23−0.34.

Experimental data were obtained by conducting four types of diagnostic tests: (1) a matching test (a task to link a prefixed verb located in the left column with a sentence in the second column); (2) a substitution test (a task to choose the missing prefixed verb (for level C1−a verbal noun) that is relevant for the corresponding context); (3) test for recognition in printed text (the task is to read the mini-text and find the prefixed verb corresponding to the given definition); (4) a test for recognition in sounding speech (the task is to listen to a mini-text and find a prefixed verb corresponding to a given definition). The first two types of tests were conducted remotely, on the educational platform we created (resource address: http://e-learningrussian.com). Recognition tests were conducted in the classroom and supplemented by corrective conversation.

The formation of the rating scale and the interpretation of the results of the diagnostic experiment were carried out in accordance with the methodology of Bespalko [30]. In his opinion, "… in the range of changes in the coefficient of assimilation of students' knowledge from 0 to 0.7, their activity has an unstable quality, they remain little sensitive to mistakes made and systematically repeat them," and "… with a coefficient of assimilation above 0.7, the activity of students acquires the necessary stability, they confidently solve problems of a given level of assimilation, are capable of self-control, have the necessary sensitivity to the mistakes they make and, as a rule, independently look for a way to correct them" [30: 61]. Thus, "the stage that provides assimilation in the K_a range from 0 to 0.7 should be called the learning process, and the stage from 0.7 to 1.0—the self-learning process" [30: 61].

Our examples are yet another proof that the verb dictionary that assists the study of the Russian language is an important tool for systematizing knowledge and should be in focus during the educational process. Compilation of a verb dictionary should be carried out taking into account such principles as the following:

(1) the principle of optimal minimization of verbal material, which is automatically formed from a verb minima for each stage of learning the Russian language;
(2) the formal lexical principle, which assumes selecting and systematizing the selected verbal units at the level of the word and phrase;
(3) the structural-semantic principle, which means the selection and systematization of the selected verbal units at the word level;
(4) the principle of historical conditioning of the formation of the verb system.

We also note that the multimedia component of the dictionary will make the process of perception of the verb by a foreigner more accurate and complete [31]. Many researchers consider computer lexicography pertaining to artificial intelligence problems.

5 Conclusion

All of the above allows us to say that the verb, as one of the complex parts of speech, requires an integrated approach for lexicographic description in different subtypes of dictionaries. A complete and accurate understanding of the meaning of a verb by a student cannot be achieved without the help of an appropriate verb dictionary that offers various ways of semantization.

At the moment, there is no resource in which verb minima would be brought into a single system. This gap can be filled by an electronic verbal dictionary, the search system of which would allow: (1) to present verb units in structural, semantic and functional aspects; (2) to show them as constituent elements of the verb system; (3) to combine the information and training potential of electronic educational resources on one platform. It is necessary to improve the functions of the educational verbal dictionary by introducing multimedia elements into the system. Such a resource will enable the user to structure and accumulate information about the verb, to make a semantic search for information in data banks and networks. In the educational process, such a resource will be helpful to a foreign user (student) in mastering programs in Russian.

References

1. Pastor, V., & Alcina, A. (2010). Search techniques in electronic dictionaries: a classification for translators. *International Journal of Lexicography—INT J LEXICOGR., 23*, 307–354. https://doi.org/10.1093/ijl/ecq015
2. Granger, S. (2012). Electronic lexicography: From challenge to opportunity. In *book: Electronic Lexicography* (pp.1–11). Publisher: Oxford University PressEditors: Sylviane Granger, Magali Paquot. https://doi.org/10.1093/acprof:oso/9780199654864.003.0001
3. Selegey D., Shavrina T., Selegey V., & Sharoff S. (2016). Automatic morphological tagging of Russian social media corpora: training and testing. Computer linguistics and intelligent technologies: based on the materials of the international conference “Dialogue 2016” Moscow
4. Alhawiti, Kh. M., & Teahan, W. J. (2014). Universal text preprocessing and postprocessing for PPM using alphabet adjustment. *Data Compression Conference, 2014*, 395–395. https://doi.org/10.1109/DCC.2014.12
5. Alhawiti, Kh. M. (2015). Advances in artificial intelligence using speech recognition. *World Academy of Science, Engineering and Technology, International Journal of Computer and Information Engineering., 9*, 1439–1442.
6. Al-Moslmi, T., Gallofré Ocaña, M., Opdahl, A. & Veres, C. (2020). Named Entity extraction for knowledge graphs: a literature overview. *IEEE Access 8*, 32862-32881. https://doi.org/10.1109/ACCESS.2020.2973928
7. Tarp, S. (2015). Preparing an online dictionary of business communication: from idea to design. Lexikos 25. https://doi.org/10.5788/25-1-1305
8. Andréka, H., Benthem, J., & Németi, I. (2017). On a new semantics for first-order predicate logic. *Journal of Philosophical Logic., 46*, 1–9. https://doi.org/10.1007/s10992-017-9429-y
9. Niccolucci, F., Hermon, S., & Doerr, M. (2015). The formal logical foundations of archaeological ontologies. In *book: Mathematics and Archaeology* (pp.86–99). https://doi.org/10.1201/b18530-5.

10. Alhawiti, K. M. (2014). Natural language processing and its use in education. *International Journal of Advanced Computer Science and Applications 5*(12), 72-76. https://doi.org/10.14569/IJACSA.2014.051210.
11. Karpov, V.I., Dobrovolsky, D.O., & Nuriev, V.A. (2019). Types of information in a bilingual dictionary. *Questions of lexicography* 16, 38-58. https://doi.org/10.17223/22274200/16/3
12. Savina, O. Y. (2016). Integration of linguistic search tools for linguodidactic purposes. Questions of lexicography 2(10), 55-66. https://doi.org/10.17223/22274200/10/4
13. Evgenieva, A.P. (Ed.). (1999). Dictionary of the Russian language: In Ch, A.P. Evgenieva (ed.) *4 volumes. RAS, Institute of Linguistic Studies*. Russian language: Polygraph resources.
14. Kuznetsova, E. V. (Ed.). (1988). *Lexico-semantic groups of Russian verbs*. Ural University Publishing House.
15. Babenko, L. G. (Ed.) (2007). A large explanatory dictionary of Russian verbs. Over 10,000 verbs. Ideographic description. Synonyms. Antonyms. English equivalents. AST-PRESS.
16. Muller, V.K. (2019). *The most complete English-Russian and Russian-English dictionary with modern transcription: about 500,000 words*. Moscow. AST.
17. Petrov, A. V. (2001). The nesting principle for the description of composites in the educational process. *Philological Studios., 2*, 130–137.
18. Sazonova, I. K. (1989). *Russian verb and its participles: Explanatory grammar dictionary*. Russian language.
19. Tikhonov, A.N. (2003). Word-formation dictionary of the Russian language: In 2 volumes: more than 145,000 words. Moscow. Astrel. AST.
20. Puritskaya, E.V., Pankov, D.I. (2018). The normative and stylistic characteristic of the vocabulary of the modern Russian language: the possibilities of description in the dictionary database. *Questions of lexicography 13*, 23-43. https://doi.org/10.17223/22274200/13/2
21. Berg, E.B., Keith, M. (2019). Finding solutions to the problems of bilingual Internet lexicography in the LexSite dictionary project. *Questions of Lexicography. 16*, 92–112. https://doi.org/10.17223/22274200/16/6
22. Barneva, R., Brimkov, V., & Stanchev, P. (2003). Children Multimedia Dictionary. International Conference Advances in Infrastructure for e-Business, e-Education, e-Science, and e-Medicine on the Internet, L'Aquila, pp. 1–6. https://www.researchgate.net/publication/2552867
23. Ziad, H., McCrae, J., & Buitelaar, P. (2018). Teanga: A linked data based platform for natural language processing. In Conference: International Conference on Language Resources and Evaluation (pp. 2410–2415). At: Miyazaki, Japan. https://www.researchgate.net/publication/325425326
24. Boonmoh, A. (2021). Use of dictionaries and online tools for reading by Thai EFL learners in a naturalistic setting. *Lexikos., 31*, 239–255. https://doi.org/10.5788/31-1-1645
25. Metruk, R. (2017). The use of electronic dictionaries for pronunciation practice by University EFL Students. *The Journal of Teaching English with Technology. 17*(4), 38–51. http://www.tewtjournal.org
26. Cimiano, Ph., Chiarcos, C., McCrae, J., & Gracia, J. (2020). Modelling lexical resources as linked data. In *Linguistic linked data* (pp. 45–59). https://doi.org/10.1007/978-3-030-30225-2_4.
27. Liang, P., & Xu, D. (2018). An empirical study of EFL learners' dictionary use in Chinese-English translation. *Lexikos., 28*, 221–244. https://doi.org/10.5788/28-1-1463
28. Andryushchenko, V.M., Morkovkin, V.V. (1988). Basic lexicographic knowledge and computer. Russian Language Machine Fund. Predesign studies. Moscow. Institute of the Russian language.
29. Atkins, B.T.S. (1996). Bilingual Dictionaries: Past, Present and Future. In M. Gellerstam et al. (eds.), (pp. 515–546). Oxford University Press.
30. Bespalko, V.P. (1989). Components of pedagogical technology. Moscow. Pedagogy.
31. Baskin, S., & Mumcu, M. (2018). Dictionary culture of university students learning English as a Foreign language in Turkey. *International Education Studies., 11*(3), 101–114. https://doi.org/10.5539/ies.v11n3p101

East Slavonic Proverbs of the “Learning—Inattention” Thematic Group (Based on the New Electronic Dictionary of Modern East Slavonic Proverbs)

Olga V. Raina, Viktoria V. Mushchinskaya, and Olga V. Guseva

Abstract The paper aims to study East Slavic proverbs of the “Learning—inattention” thematic group. These proverbs reflect the national paremiological representation of the world, pertaining the cultural identity of the countries in question. Thoughts about the benefits of knowledge are the most important part of proverbial folklore. A comparison of proverbs from different peoples shows how much these peoples have in common, which, in turn, contributes to their better mutual understanding and rapprochement. There are international units that exist in all languages. They are from ancient Greek, Latin and Biblical texts. The majority of Russian proverbs have equivalents in Belarusian and Ukrainian. The material collected for the new *Electronic Dictionary of Modern East Slavonic Proverbs* indicates that the developed lexicographic concept will allow to present the East Slavic paremiological space in the unity and diversity of structural and semantic models, language images, and will contribute to the improvement of the system of comparative paremiography.

Keywords Proverb · Proverbial parallel · Russian language · Belarusian language · Ukrainian Language · Sociolinguistic paremiological experiment · Electronic dictionary

O. V. Raina (✉) · V. V. Mushchinskaya · O. V. Guseva
St. Petersburg University, Universitetskaya Nab., 7-9, St.Petersburg 199034, Russia
e-mail: o.raina@spbu.ru

V. V. Mushchinskaya
e-mail: v.mushinskaya@spbu.ru

O. V. Guseva
e-mail: o.guseva@spbu.ru

P. Eismont et al. (eds.), *Literature, Language and Computing*,
https://doi.org/10.1007/978-981-99-3604-5_13

1 Introduction

Paremiology allows you to obtain information about the representation of the world of a certain nation. Proverbs, being a storehouse of experience accumulated over the centuries, contain important information not only about current values, but also about already forgotten times, thus representing an incredible linguistic and cultural significance. Researchers became interested in Russian-Belarusian-Ukrainian paremiology no so long ago. *The Big Russian-Belarusian Dictionary of Proverbs* in two parts was published in 2001. Z. Sanko published the *Small Russian-Belarusian Dictionary of Proverbs* in 1991. The first collections of Ukrainian proverbs of the nineteenth century were traditionally compiled in alphabetical order (the collections by K. Zinoviev, A. Pavlovsky, V. Smirnitsky, G. Ilkevich, N. Zakrevskii, and others). Fundamental works in the field of Ukrainian paremiography include the M. Nomis's collection *Ukrainian Proverbs* (1901–1910) and a three-volume edition of Ukrainian folk wisdom by M. M. Pazyak (Kyiv, 1989–1991). Modern Ukrainian paremiographers not only reprint the collections of proverbs, but also develop new types of dictionaries (T. Kosmeda, A. Pedersen, K. Serova, etc.). This study is aimed at defining the Belarusian and Ukrainian proverbial parallels of Russian proverbs that will be included in the *Electronic Dictionary of Modern East Slavonic Proverbs (*also called *Dictionary of Current Active East Slavonic Proverbs*), which is currently being compiled at the Department of Slavonic Philology of St. Petersburg State University (by the authors of this paper, as well as by M. Yu. Kotova, N. E. Boeva, and O. S. Sergienko). This dictionary demonstrates currently used Belarusian and Ukrainian proverbial parallels of the Russian paremiological minimum. *The Electronic Dictionary* is based on the proverbs included in the *Russian-Slavonic Dictionary of Proverbs with English Parallels*, published in 2000 by M. Yu. Kotova [1]. A team of researchers led by Prof. M. Yu. Kotova is engaged in identifying and studying the parallels of the Russian paremiological minimum in each of the Slavonic languages presented in the dictionary [2–5].

2 Purpose of the Study

This research will display the East Slavonic paremiological core, which reflects the common segment of Russian, Belarusian and Ukrainian proverbial representations of the world, as well as lacunae in the Belarusian and Ukrainian paremiological corpora in relation to Russian proverbs of the thematic group "Learning—inattention". All the results of this research will be presented in the new *Electronic Dictionary of Modern East Slavonic Proverbs.* The main concept of comparative electronic dictionary representation of East Slavic proverbs developed by the authors of the article, as well as M. Yu. Kotova, N. E. Boeva, O. S. Sergienko is presented. The novelty of the study lies in the analysis of paremiological data that has not been developed in a comparative lexicographical aspect and an innovative approach to its lexicography. It is proved that the value of the electronic dictionary as an ethno-linguistic source will increase due to a combination of thematic organization of the material, marking similarities and differences when presenting paremiological parallels. Particular attention is paid to the identification and vocabulary representation of full paremiological equivalents, the establishment of correspondence types for partial equivalents, and the description of their ethno-specific elements.

3 Research Methods

The methodological grounds of the research are the system-structural paremiographic approach and the method of the comparative cultural linguistics studies. The authors of the dictionary have conducted sociolinguistic paremiological experiments among na- tive speakers of the Russian, Belarusian and Ukrainian languages, defining the frequency of usage of modern Belarusian and Ukrainian proverbial parallels today [6]. The study to identify the parallels of the Russian paremiological minimum involved all East Slavic languages presented in the dictionary. The results of the sociolinguistic paremiological experiment were sequentially processed; each proverb indicated by the informants was checked in the Internet for usage, supplemented with contextual illustrations or notes about their absence [7]. Each Belarusian proverbial parallel has been tested in the following sources: *Handbooks of a Paremiorapher. Issue 5. Belarusian Proverbial Parallels of Russian Proverbs from Paremiological Minimum* [2], corpus (Belarusian) [8], the national corpus of the Belarusian language [9], the dictionaries of Belarusian proverbs by I. I. Nosovich [10], Ya. N. Rapanovich [11] and Z. Sanko [12], *the Russian-Belarusian Dictionary of Proverbs* by E. E. Ivanov [13], and the *Etymological Dictionary of Proverbs* by I. Ya. Lepeshev [14]. Each Ukrainian proverbial parallel has been tested in the corpus of the Ukrainian language [15], the dictionary by M. M. Pazyak [16], *the Russian-Ukrainian Dictionary of Proverbs* [17], *the Dictionary of Russian Proverbs and Sayings with Ukrainian Equivalents* [18], *the Etymological Dictionary of the Ukrainian Language* [19], *the Phraseological Dictionary of the Ukrainian Language* [20]. Lacunae in Belarusian or Ukrainian paremiological corpora are represented in the dictionary as a literal translation of the Russian proverbs into Belarusian and Ukrainian.

4 Results

The Electronic Dictionary of Modern East Slavonic Proverbs has thematic index of modern Russian proverbs introduced in the 1st part. Let us examine here the Belarusian and Ukrainian proverbial parallels of the currently used Russian proverbs in the “Learning, knowledge—ignorance, inattention” thematic group. This list contains 17 Russian proverbs. All definitions of the Russian proverbs [7] are translated by the authors of the article from *the Electronic Dictionary of Modern East Slavonic Proverbs.*

Russian *Vek zhivi—vek uchis'*. cf. full English proverbial parallel *Live and learn* [21: 97]. The knowledge that people receive in their youth is not enough, because throughout life they are convinced of how much they do not know in order to improve their life. The proverb says that a person must be open to new knowledge all his/her life. For the first time this catchphrase was used by the Roman Stoic philosopher Lucius Annaeus Seneca, also known as Seneca the Younger in the first century AD: "Live a century—learn a century how to live". In the following centuries, the aphorism was quoted many times and lost its ending. From Latin literature, the saying was borrowed by many European languages. As for the Russian language, the proverb has a long history. It is recorded in *the Dictionary of the Russian Language of the 18th Century*; in the nineteenth century it was included into the dictionary of I. M. Snegirev (1831, 1848), while in the twentieth century it was added to the dictionaries of proverbs by A. I. Sobolev (1956), A. S. Spirin (1985), A. A. Razumova (1957), and others. M. I. Mikhelson in his dictionary cites Latin as a source of borrowing of the Russian proverb: *Tamdiu discendum est, quamdiu nescias: quamdiu vivis.* In V. Dahl's book *Proverbs of Russian People* (1853), this proverb is found in an expanded version: *Vek zhivi—vek uchis' (a umri durakom)*/lit. '*Live a century—learn a century [and die a fool]'/.* The ironic proverb *you will die a fool* speaks of the futility of teaching: no matter how much you study, you still cannot learn everything. The proverb is well known in Ukrainian: *Vik zhyvy, vik uchys*, as well as in Belarusian: *Vek zhyvi, vek vuchysya.* In *the Russian-Belarusian Dictionary of Proverbs* by E. E. Ivanov (2001), several Belarusian equivalents are given: *Buzdem vuchytstsa, pakul' smerts' luchytstsa (naluchytstsa)* /lit. '*We will learn until death unites (joins)*'/, *Vek zhyvem, vek vuchymsya*/lit. '*We live a century, we learn a century*/, *Vek zhyvesh, vek vuchyshsya*/lit. '*You live a century, you learn a century*'/, *Shto zhyvem, to vuchymsya*/ lit. '*As we live, we learn*'/, *Shto zhyvesh, shto vuchyshsya*/lit. '*As you live, you learn'*/ , *Vek zhyvi <i> vek uchysya, a use dernem pamrehsh*/lit. '*You live a century <and> learn a century, and you'll still die like a fool*'/[6: 35- 36]. In the *Electronic Dictionary of Modern East Slavonic Proverbs* we find Belurasian *Vek zhyvi, vek vuchysya* and Ukrainian *Vik zhyvy, vik uchys* proverbial parallels. They are full equivalents (the same image is used, as well as the same component composition).

Russian *Cheloveku (Lyudyam) svojstvenno oshibat'sya,* cf. English proverbial par- allel *To err is human.* Spoken with a desire to reassure someone who has made a mis- take. The proverb says that it is necessary to treat indulgently human weaknesses, mis- takes and imperfection. It is rooted in ancient literature. The Greek poet Theognis said: "Mistakes are inevitable among mortals". Euripides echoed him: "All people tend to make mistakes." The famous maxim *To err is human* (lat. *Errare humanum est*) is attributed to the Roman writer-rhetorician Mark Annaeus Seneca (c. 55 BC–c. 37 AD), father of the famous philosopher Seneca the Younger. The ancient Roman politician and orator Mark Thulius Cicero (106–43 BC) supplemented this phrase*: Each person tends to err, but only fools repeat their mistakes.* The phrase has become a catchphrase widely used in European languages. It is often found in fiction. Mikhelson's phraseological dictionary illustrates the proverb *To err is human* with examples from Turgenev, Saltykov-Shchedrin, as well as from other European authors: J. W. von Göthe, A. Pope and etc. The proverb is also well known in East Slavonic languages. We find this proof in *the Electronic Dictionary of Modern East Slavonic Proverbs*: Belarusian *Chalaveku (Lyudzyam) ulastsiva pamylyatstsa*, Ukrainian *Liudyni vlastyvo pomyliatysia.* They are full equivalents (the same image is used, as well as the same component composition).

Russian *Na oshibkah uchatsya.* /lit. ‘*Learn from mistakes*’/, cf. English proverbial parallels *Adversity makes a man wise, not rich* [21:23]; *Failure teaches success* [5:56]; *Sorrow makes us wise* [21:132]. Encouragement to further action to one who repents of a mistake. The proverb says that you need to learn from your mistakes, to recognize that mistakes inevitably accompany any activity. The proverb was included in the dic- tionaries of proverbs only in the twentieth century: it is recorded in the dictionaries by M. A. Rybnikova (1961), A. S. Spirin (1985), V. Tanchuk (1986). Perhaps it came from a proverb indicated in the book *Proverbs of Russian people* (1853) by V. I. Dahl (section Oversight–Quickness): *The one who does nothing is not mistaken; and whoever does is wrong.* But it can also be assumed that this is a development of the Latin maxim *Errando discimus.* There are full equivalents (the same image is used, as well as the same component composition) in Belarusian *Na pamylkakh vuchatstsa* and Ukrainian *Na pomylkakh vchatsia /rozumni liudy—na chuzhykh/.*

Russian *Ne oshibaetsya tol’ko tot, kto nichego ne delaet* /lit. ‘*Only he who does nothing does not make mistakes*’/, cf. English proverbial parallel *He that never climbed never fell* [21:74]. Consolation to people who are worried that they made a mistake, encouraging them to take further action. The popular proverb means that it is impossible to do without mistakes in work, but there is no need to be afraid of them. V. I. Dahl in the book *Proverbs of Russian people* (1853) cites the proverb *The one who does nothing is not mistaken; and whoever does is mistaken.* This version appeared in the *Collection of Figurative Words and Allegories* (1904) fifty years later. Mikhelson in his *Big Explanatory Phraseological Dictionary* presents the maxim of the ancient Greek playwright Euripides (480–406 BC) *A person who does a lot, and is wrong in many ways* as illustration of this proverb. Despite such a long history, the aphorism *The one who does nothing is never wrong* is attributed to both the 26th US President Theodore Roosevelt (1858–1919) and V. I. Lenin (1870–1924), who said at the 5th Congress of Soviets on July 5, 1918: “The one who does nothing practical is not mistaken.” The proverb is known and currently used in other East Slavonic lan- guages. They are full equivalents (the same image is used, as well as the same compo- nent composition): Belarusian *Ne pamylyaetstsa toj, khto nichoga ne robits’* and Ukrainian *Ne pomyliaietsia toi, khto nichoho ne robyt.*

Russian *Povtoren’e—mat’ uchen’ya* /lit. ‘*Repetition is the mother of learning’*/, cf. English proverbial parallels *Experience is the mother of wisdom* [21:55]; *Practice makes perfect* [21:125]; *Use makes the craftsman* [21:167]. The proverb means that deep knowledge is gained only by those who constantly return to what was previously studied and revise the material. The Latin maxim underlying the Russian proverb is given by M. I. Mikhelson*: Repetitio est mater studiorum.* The author of the Latin maxim was the Roman poet Publius Ovidius Naso, the full version of the quote sounds like *Repetition is the mother of learning and the refuge of donkeys (consolation of fools).* This was the author’s commentary on the traditional education system, in which only not very gifted students memorized material by repetition. From Latin the aphorism came to other European languages: German *Wiederholen heisst lernen*, French *Répéter c'est reculer pour mieux sauter*. This proverb is found in the book *Proverbs of Russian people* (1853) by V. I. Dahl (section Study—Science), as well as in the dictionaries of proverbs of the twentieth century. The ironic version of the proverb was born in the twentieth century as well: *Povtoren'e mat' uchen'ya, no i pribezhishche dlya lentyaev* /lit. '*Repetition is the mother of learning, but also a refuge for lazy people'*/, which refers us to the original Ovid's maxim. The proverb is also known and currently used in other East Slavonic languages. There are full equivalents (the same image is used, as well as the same component composition) in Belarusian *Pautarehnnya matsi vuchehnnya* and Ukrainian *Povtorennia—maty navchannia.*

Russian *Uchen'e—svet, neuchen'e—t'ma* /lit. '*Learning—light, ignorance—dark- ness*'/, cf. English proverbial parallels *Learning is the eye of the mind*; *Learning makes wise, ignorance otherwise* [21:92]. It is necessary to study in order to be a well-educated person. This proverb first appeared in the collection of proverbs by V. N. Tatishchev in the beginning of the eighteenth century. It is given in the same version by V. I. Dahl in the book *Proverbs of Russian people* (1853) and M. I. Mikhelson in the *Big Explanatory Phraseological Dictionary* (1904), as well as in the dictionaries of proverbs of the twentieth century. The aphoristic structure of the proverb led to the appearance of its modern translation into Latin: *Scientia nihil aliud est quam veritas*, although the proverb was not used in Ancient Rome. The twentieth century saw the birth of ironic versions of the proverb that have a similar structure: *Uchen'e svet, a neuchen'e—chut' svet i na rabotu* /lit. *'Learning is light and ignorance is to go to work at the crack of dawn*'/, *Uchen'e svet, a neuchenyh t'ma* /lit. *'Learning is light, but there are thousands of unlearned'/.* There is a relative equivalent (incomplete coincidence of structure and complete coincidence of figurativeness) in the Ukrainian language *Uchenomu—svit, nevchenomu—tma* /lit. '*To the educated—light, to the ignorant—darkness*'/. In the Belarusian language, both the full equivalent of the Russian proverb *Vuchehnne svyatlo, a nevuchonym tsemra*, as well as the analogue (different images and different component composition) *Bez navuk yak bez ruk* /lit. *'Without science/education you are helpless/without hands*'/ are used.

Russian *Chto napisano perom—ne vyrubish' toporom* /lit. '*What is written with a quill—you can't cut it down with an ax*'/. You should better think over the content and form of the words you write, as you will not be able to change them later. The proverb means that if what is written has become known, then you cannot change or correct it. The origin of the proverb comes from a respectful attitude towards official documents in Russia. People were to obey the orders given by the prince implicitly. Another version of the origin of the proverb is associated with parchment, the material that was used in ancient times for writing. The ink soaked into the finely crafted skin of parchment so deep that it had to be scraped off with a knife or pumice to reuse it, and traces still remained. Centuries later, scholars can read ancient palimpsests. M. I. Mikhelson in his *Big Explanatory Phraseological Dictionary* (1904) offers a Latin maxim to which the Russian proverb can be traced: *Verba volant, scripta manent.*

The Russian proverb is recorded in the dictionaries of the eighteenth century: in the collection of proverbs by V. N. Tatishchev and in the handwritten *Collection of Proverbs of the Former Petrovsky Gallery*. In the book *Proverbs of Russian people* (1853) by V. I. Dahl there are variants of the proverb: *Napishesh' perom, ne steshesh' (ne vyrubish') toporom* /lit. '*You will write with a quill, you won't trim (cut it out) with an ax*'/, *Napisano perom, ne vyrubit' i toporom* /lit. '*It's written with a quill, you can't cut it down with an ax*'/, *Napishesh' perom, chto ne vyvezesh' volom* /lit. '*You will write with a quill and you can't pull it out with an ox*'/. In the Belarusian language there is a full equivalent (the same image is used, as well as the same component composition) *Shto napisana pyarom, ne vyrubish* syakeraj.* A feature of Russian paremia is the internal rhyme, which is lost in Belarusian. In Ukrainian, we also find both the relative equivalent (inconsistency in structure and coincidence in figurativeness) *Napysanoho sokyroiu ne vyrubaiesh* and the analogue of the Russian proverb (different structure and figurativeness) *Napysanoho perom ne vytiahnesh i volom* /lit. '*You can't pull out by an ox what is written with a quill*'/.

Russian *A Vas'ka slushaet da est* /lit. '*And Vaska is listening and eating*'/. (1) About headstrong children who do not pay attention to adults' comments and get their way. (2) About adults who, despite criticism, continue doing something, rather impassively. This proverb has a literary origin. It came into the language from the fable by I. A. Krylov *The Cat and the Cook* (1813). The fable tells the story of a cook who, having caught a cat eating chicken, began to lecture him on morals. The cat listened inattentively and continued to eat. The moral of the fable is when it is possible and necessary to use force, one should not waste time on idle chatter. The fable appeared during the Patriotic War of 1812 and expressed the political views of the author. Krylov condemned the policy of Emperor Alexander I, who, even after the French troops crossed the Russian borders, still hoped for a peaceful solution to the conflict and sent a letter to Napoleon urging him to end the war and avoid bloodshed. The cat and the cook are images of two emperors. Despite the fact that Krylov's fables instantly became sources of catchphrases, this proverb was recorded in the phraseological dictionary by M. I. Mikhelson only in the nineteenth century. It was later included in the dictionaries of Russian proverbs of the twentieth century. Due to its literary origin, this paremia has no equivalents in other languages. But common history, as well as a common syllabus of the course on Russian literature in Soviet schools, led to the fact that loan translations from Russian appeared in the Belarusian and Ukrainian languages: Belarusian full equivalent *A Vas'ka slukhae dy ests'* and Ukrainian relative equivalent (figurative base is similiar, but the component composition has differences) *A muryi yist sobi ta yist* /lit. '*And the wise one eats and eats'*/.

Russian *Bumaga vse sterpit* /lit. *Paper will endure everything* '/, cf. English proverbial parallel *Youth and white paper take any impression.* It is about mediocre writing or about the egregious facts that are stated in the documents. The proverb came to European languages from Latin. Its primary source is the message "To Friends" of the Roman statesman, orator and writer Marcus Thullius Cicero (106–43 BC), who wrote that *Epistola non erubescit* /lit. '*Paper does not blush*'/. In Russian, both loan translation from Latin is known *Bumaga ne krasneet*, and the allegorical interpretation of the aphorism: you can write anything you want, although not everything that is written is true. Mikhelson offers variants of proverbs: *Bumaga vse terpit (ne krasneet) / ‘Paper endures everything (does not blush)’/.* Perhaps the expression came to Russian from French, in which both options are found: *Le papier endure tout* or *Le papier souffre tout*/lit. ‘*Paper will endure/ endure everything*’/, and *Le papier ne rougit pas* /lit. ‘*Paper does not blush*’/. A variant of the proverb *Bumaga vse sterpit* is recorded in the *Dictionary of the Russian Language of the 18th Century.* The popular expression of Cicero served as a source of loan translation in Belarusian *Papera use stserpits’* and Ukrainian *Papir vse vytrymaie.* They are full equivalents (the same image is used, as well as the same component composition).

Russian *V odno uho voshlo, /a/ v drugoe vyshlo* /lit. ‘*It went in one ear, / and / went out the other*’/. About an adult or a naughty child who does not respond to critical remarks, continuing to do what he/she is scolded for (cf. Russian proverb *A Vas’ka slushaet da est*). You can find it in the dictionary by Mikhelson with the meaning ‘as you heard—immediately forgot’, about the one who listens without attention and about the one who soon forgets. The paremia was included in the *Big Dictionary of Russian Sayings* (2007), ed. by V.M. Mokienko. Belarusian proverbial parallel *U adno vukha ŷlyatsela, u drugoe—vyletsela* /lit. ‘*It flew into one ear and out the other*’/ is the relative equivalent (inconsistency in structure and coincidence in figurativeness). In Ukrainian we also find the relative equivalent *V odne vukho vkhodyt (vletyt), v inshe (chere*z̦ *inshe,* z̦ *inshoho)—vykhodyt (vyletyt)*/lit. ‘*It goes (flies) in one ear, it goes out (flies out) the other*’/.

Russian *Mnogo budesh’* z̦*nat’—skoro sostarish’sya* /lit. ‘*If you know a lot, you will soon grow old*’/, cf. English proverbial parallels *Too much knowledge makes the head bald* [21:163]. Whoever seeks to acquire a large amount of knowledge, runs the risk of quickly losing the carelessness and naivety inherent in youth (it is said: a) by elders to younger people or b) ironically in a conversation between peers). This proverb is included in the book *Proverbs of Russian people* (1853) by V. I. Dahl. It is also given in the *Explanatory Dictionary of the Great Russian Language* as an illustration of the word *know*. Mikhelson refers to the Latin proverb *Is cadet ante senem, qui sapit ante diem* /lit. ‘*He who dies before he grows old knows (a lot)*’/. In Russia adults often recall the proverb ironically and jokingly when, for some reason, they do not want to explain something to children. The proverb has equivalents in Belarusian *SHmat bud*z̦*esh vedats’—khutka (skora) pastarehesh (sastaryshsya)* and Ukrainian *Bahato* z̦*natymesh, shvydko postariiesh.*

Russian *Ne uchi uchenogo* /lit. ‘*Don’t teach a scholar/smb who’s educated’*/, cf. English proverbial parallel *Don’t teach your grandmother to suck eggs; Don’t teach the dog to bark* [21:47]; *Old foxes want no tutors* [21:116]. (rude) Don’t try to advise someone who knows this subject better than you. This proverb came into Russian from the comedy by the Roman comedian T. M. Plautus. Since a Latin literary source made this paremia international, its translation can be found in different languages, including Ukrainian and Belarusian. There are full equivalents in Belarusian *Ne vuchy vuchonaga* and Ukrainian *Ne vchy uchenoho.* The author of the Belarusian

Etymological Dictionary of Proverbs I. Ya. Lepeshev writes that the Belarusian proverb was borrowed from the Plautus’s comedy, citing its Latin original: *Omnem operam perdis—Quid doctum doces* /lit. ‘*You waste your efforts teaching a scholar*’/ [14: 86].

Russian *Slona-to on i ne primetil* /lit. ‘*He did not notice the elephant*’/. About an inattentive and not very knowledgeable person who pays attention to the secondary thing and does not notice the main one. It goes back to I. A. Krylov’s fable *The Curious Man* (1814). According to the plot, a visitor to the museum was carried away by looking at small insects there, but, as it turns out, he did not notice something that was impossible to be missed, namely an elephant. Mikhelson cites a variant of the proverb *I didn’t even notice the elephant* with the meaning ‘*the main thing was overlooked*’. *The Big Dictionary of Russian Sayings* (2007) cites the proverb *Slona ne primetit’*. The presence of full equivalents in Belarusian *Slana-to yon i ne zauvazhyu* and Ukrainian *Slona ne pomityly* can be explained by a common history and syllabus of the course on Russian literature in Soviet schools.

Russian *Slushaj uhom, a ne bryuhom* /lit. ‘*Listen with your ear, not your belly*’/. (rude) Advice to be attentive to explanations and instructions, and not to scorn them. The paremia was recorded in the *Collection of Proverbs* by A. I. Bogdanov (1741) for the first time. It is found in the Dahl’s dictionary, as well as in the dictionaries of proverbs of the twentieth century. In Belarusian we find the full equivalent *Slukhaj vukham, a ne brukham.* This proverb can be found in the Nosovich’s dictionary as well [8: 151]. There is an analogue (different images and different component composition) in Ukrainian *Slukhai oboma vukhamy* /lit. ‘*Listen with both ears*’/.

Russian *Slyshal zvon, da ne znaet, gde on* /lit. ‘*Heard the bell ringing, but does not know where it is*’/. About a person who has information about something, but does not know how to use it. This proverb is used disapprovingly of a person who knows something only partially, vaguely, therefore he/she is mistaken, speaks out of place, malapropos. In the old days, people often focused on the church bell ringing, because they told the time with its help, and sometimes could identify the location of the church with the bell tower. The ringing of bells resounded far, sometimes it was impossible to understand where it came from, or to misinterpret it. This paremia was recorded in Mikhelson’s phraseological dictionary for the first time. A feature of the proverb is internal rhyme. There are full equivalents in Belarusian *Chue zvon, dy ne vedae dze (adkul’) (yon)*and Ukrainian *Chuie dzvin, ta ne znaie, de vin.*

Russian *Smotrit v knigu, a vidit figu* /lit. ‘*He looks into the book, but sees nothing/ the fig sign (gesture used to deny a request)*’/. (1) About a poorly educated person who reads, but does not understand anything from the text. (2) About a person who seems to be immersed in reading, but does not understand a single line of what he/she read because of daydreaming. It was recorded in the dictionary of proverbs by V. I. Dahl for the first time. The peculiarity of this proverb is internal rhyme. The paremia is found in the twentieth century dictionaries by A. A. Razumov (1957) and by V. P. Anikin (1988); it is also included in the *Big Dictionary of Russian Proverbs* by V. M. Mokienko. There are full equivalents in Belarusian *Glyadzits’ u knigu, bachyts’ figu* and in Ukrainian *Dyvytsia v knyhu, a bachyt fihu.*

Russian *Tebe russkim yazykom govoryat /, a ty ne ponimaesh'/!* /lit. '*They/I speak to you in Russian/but you don't understand /!'*/. The proverb is used to speak with irritation about those who do not respond to clear and obvious words. It is included in the M. I. Mikhelson's dictionary with the meaning: clearly, understandably, truly. The paremia is formed according to the structural-semantic model known in different European languages since antiquity. M. I. Michelson gives such examples from European languages: *Deutsch mit einem sprechen*, *Parler français; Latin loqui* [22]. The clear association with the Russian language determines the lack of proverbial equivalents in other languages.

5 Conclusion

This paper considers the modern East Slavonic parallels of the Russian proverbs of the "Learning, knowledge–ignorance, inattention" thematic group which were identified in the course of the sociolinguistic paremiological experiments included into the *Electronic Dictionary of Modern East Slavonic Proverbs.*

14 Russian proverbs have equivalents, while one has a relative equivalent in Belarusian. One Russian proverb *Uchen'e—svet, neuchen'e—t'ma* has both Belarusian equivalent and analogue.

11 Russian proverbs have equivalents, while three—a relative equivalent and one—an analogue in Ukrainian. One Russian proverb *Chto napisano perom—ne vyrubish' toporom* has both Ukrainian relative equivalent and analogue.

The presence of a large number of Belarusian and Ukrainian equivalents is explained by the common history and school curriculum.

There are 8 international units that exist in different languages, including non-Slavic ones, and have ancient Greek, Latin, and Biblical texts as their source. Such units can penetrate different languages independently of each other.

The paremia *Tebe russkim yazykom govoryat /, a ty ne ponimaesh'/!* is a single lacuna for Belarusian and Ukrainian that can be explained by its national peculiarity.

The proverbs are not only a source of worldly wisdom, but also answers to eternal questions that concern many generations of people. Knowledge for a person is a kind of guide that allows one to navigate in the world and helps to comprehend it. However, knowledge is multifaceted and unpredictable, and it can become a terrible weapon in the hands of an ignorant person. The eternal confrontation between knowledge and ignorance, learning and inattention is an inexhaustible source for reflection.

Acknowledgements This paper is written within the framework of the project *Electronic Dictionary of Modern East Slavonic Proverbs* (20-012-00105/22) funded by the Russian Foundation for Basic Research (RFBR).

References

1. Kotova, M. Y. (2000). Russko-slavyanskiy slovar poslovits s angliyskimi sootvetstviyami. St. Petersburg: Izdatelstvo Sankt-Peterburgskogo universiteta, 2000 [Russian-Slavonic Dictionary of Proverbs with English Parallels]. St. Petersburg: Saint Petersburg State University (In Russian)
2. Kotova, M. Y., Boeva, N. E. (2019). Tetradi paremiographa. Vypusk 5: Belorusskie poslovichnyje paralleli russkih poslovits paremiologicheskogo minimuma. St. Petersburg: Filologicheskiy fakultet SpbGU [Handbooks of a Paremiorapher. Issue 5. Belorussian Proverbial Parallels of Russian Proverbs from Paremiological Minimum. M. Y. Kotova (ed.)] St. Petersburg: VVM Publ. 304 p.. (In Russian)
3. Kotova, M. Y., Raina, O. V. (2020). Towards a linguistic vision of the world at the paremiological level of language. *Vestnik of Saint Petersburg University. Language and Literature, 17*(3), 487–504
4. Kotova, M. Y., Mushchinskaya, V. V., Sergienko, O. S. (2020). East Slavonic proverbs with ethnonyms in the electronic dictionary of current proverbs. In: *The European Proceedings of Social and Behavioral Sciences EpSBS. WUT 2020. 10th International Conference “Word, Utterance, Text: Cognitive, Pragmatic and Cultural Aspects”* (pp. 748–757)
5. Mushchinskaya, V. V. (2021). Problematika podbora poslovichnyh ukrainskih parallelej v Elektronom slovare sovremennyh aktivnyh vostochnoslavyanskih poslovic [Problems of selection of proverbial Ukrainian parallels in the Electronic Dictionary of Modern East Slavonic Proverbs]. In: Yazyk. Kul’tura. Kommunikaciya: izuchenie i obuchenie. Sbornik nauchnyh trudov V Mezhdunarodnoj nauchno-prakticheskoj konferencii (14–15 oktyabrya 2021 g., g. Orel, OGU imeni I.S. Turgeneva). – Orel, OGU imeni I.S. Turgeneva, izdatel’stvo «Kartush» (pp. 519–524). (In Russian)
6. Kotova, M. Y.: (2021) O sociolingvisticheskom paremiologicheskom eksperimente s uchastiem nositelej ukrainskogo yazyka [On the sociolinguistic paremiological experiment with the participation of native speakers of the Ukrainian language]. In: Sbornik materialov Mezhdunarodnoj konferencii po estestvennym i gumanitarnym naukam Sankt-Peterburgskij gosudarstvennyj universitet «Science SPbU – 2020». SPb.: Izdatel’stvo: OOO “Skifiya-print” (pp. 1424–1425). ISBN 978-5-98620-509-0 (In Russian)
7. Kotova, M. Y., & Sergienko, O. S. (2021). Principy paremiografirovaniya v Elektronnom slovare sovremennyh aktivnyh vostochnoslavyanskih poslovic (na materiale russko-belorusskih poslovichnyh parallelej) [Paremiographic principles in the Electronic Dictionary of Modern East Slavonic Proverbs (As seen in the Russian-Belorussian paremiological parallels)]. *Slavica Slovaca, 56*(2), 242–251. (In Russian).
8. Nacional’nyj korpus russkogo yazyka. Parallel’nyj korpus [National corpus of the Russian language. Parallel corpus]. Retrieved 18 August, 2021, from https://ruscorpora.ru/new/search-para-be.html (In Russian)
9. Belaruski N-korpus [Belarusian N-corpus]. Retrieved 18 August, 2021, from https://bnkorpus.info/index.html (In Belarusian)
10. Nosovich, I. I. (1874). Sbornik belorusskih poslovic [Collection of Belarusian Proverbs]. SPb.: V tip. Imp. Akademii nauk, 232 p. (In Russian)
11. Rapanovich, Y. N. (1974) Belaruskiya prykazki, prymawki i zagadki [Belarusian proverbs, sayings and riddles]. Minsk: Vyshejshaya shkola, 384 p. (In Belarusian)
12. San’ko, Z. (1991). Maly ruska-belaruski slownik prykazak, prymavak i frazem [Small Russian-Belarusian dictionary of proverbs, sayings and phrasems]. Minsk: Navuka i tekhnika (58 p.) (In Belarusian)
13. Ivanov, E. E. (2001). Russko-belorusskij slovar’ poslovic = Ruska-belaruski sloŷnik prykazak: 777 poslovic russkogo yazyka, svyshe 5000 belorusskih paremiologicheskih ekvivalentov i sootvetstvij [Russian-Belarusian Dictionary of Proverbs: 777 proverbs of the Russian language, over 5000 Belarusian paremiological equivalents and correspondences]. Mogilev: GA MT «Brama» (Vol. 1–2) (In Russian and Belarusian)

14. Lepeshaў, I. Y. (2014). Etymalagichny slownik prykazak [Etymological Dictionary of Proverbs]. Minsk: Vyshejshaya shkola, 141 p. (In Belarusian)
15. Korpus ukrainskoi movy. Retrieved 18 August, 2021, from http://korpus.org.ua/ (in Ukrainian)
16. Paziak, M. M. (2001). Ukrainski pryslivia, prykazky ta porivniannia z literaturnykh pamiatok. Kyiv.: Naukova dumka (In Ukrainian)
17. Praktychnyi rosiisko-ukrainskyi slovnyk prykazok [Practical Russian-Ukrainian Dictionary of Proverbs] (2009). Uporiad. H. Mlodzynskyi; Za red.M. Yohansena. Derzhavne vydavnytstvo Uk rainy, Kyev (In Ukrainian)
18. Rosiiski pryslivia ta prykazky z ukrainskymy vidpovidnykamy [Russian Proverbs and Sayings with Ukrainian Equivalents] (1969). Vydannia druhe, vypravlene y dopovnene. Ukladachka N. Bielenkova. Kyiv: Vydavnytstvo khudozhnoi literatury «Dnipro» (In Ukrainian)
19. Etymolohichnyi slovnyk ukrainskoi movy: v 7 t. [Etymological Dictionary of the Ukrainian Language: In 7 vols.] K.: Nauk. dumka (1982–2006). (In Ukrainian)
20. Frazeolohichnyi slovnyk ukrainskoi movy [Phraseological Dictionary of the Ukrainian Language]. Kyiv: Nauk. dumka (1993). (In Ukrainian)
21. Kotova, M. Y., Kolpakova, A. A. (2018). Tetradi paremiographa. In M. Y. Kotova (ed.) Vypusk 4: Anglijskie poslovichnyje paralleli russkih poslovits paremiologicheskogo minimuma (na fone bolgarskih, slovatskih i czeshskih poslovits) [Handbooks of a Paremiorapher. Issue 4. English Proverbial Parallels of Russian Proverbs from Paremiological Minimum (with the corresponding Bulgarian, Slovak and Czech proverbs)] (188 p.). St. Petersburg: Contrast Publ., (In Russian)
22. Mikhelson, M. I. (1896–1912). Russkaya mysl' i rech'. Svoe i chuzhoe. Opyt russkoj frazeologii. Sbornik obraznyh slov i inoskazanij. T. 1—2. Hodyachie i metkie slova. Sbornik russkih i inostrannyh citat, poslovic, pogovorok, poslovichnyh vyrazhenij i inoskazanij [Russian thought and speech. Ours and someone else's. Experience of Russian phraseology. Collection of figurative words and parables. V. 1–2. Sayings. Collection of Russian and foreign quotations, proverbs, sayings, proverbial expressions and allegories]. SPb., tip. Ak. nauk. Retrieved 18 August 2021, from https://gufo.me/dict/mikhelson (In Russian)

Online Peer-To-Peer Teaching Practice as a Tool to Advance Professional Competence in Distant Learning Format

Olga Antciferova, Tatyana Kolosova, and Kira Shchukina

Abstract The article describes peer-to-peer online practice as a new format of online teaching. The aim of the research is to develop this concept to define its linguistic and methodology value for contemporary educational environment and to identify additional professional skills, enabled through peer-to-peer online practice. Methodology basis of the research includes works on cooperative learning, e-learning, linguodidactics, as well as academic papers on digital technologies integration with methods used to teach Russian as a foreign language. We also rely on recent developments in e-learning methodology, online resources and tools for teachers. Identifying, describing and implementing a new type of teaching practice links independent learning and authentic interaction using a foreign language to promote professional satisfaction among future educators, working in cross-cultural information environment. Individualized teaching technology relying on online resources justifies the theoretical relevance of our research. Peer-to-peer online practice is teacher–student interaction, allowing to adapt knowledge acquisition to the student's individual psychology and particular practical needs, while maintaining standard methodology of teaching. This type of interaction (peer-to-peer) hasn't been described in detail among contemporary methods of teaching Russian as a foreign language, especially as a tool to train future specialists.

Keywords Professional competence · Cooperative learning · Peer-to-peer learning · Peer-to-peer online practice · E-Linguodidactics in teaching Russian as a foreign language

O. Antciferova · T. Kolosova · K. Shchukina (✉)
St. Petersburg University, Universitetskaya Emb. 7/9, 199034 St. Petersburg, Russia
e-mail: k.shukina@spbu.ru

O. Antciferova
e-mail: o.antciferova@spbu.ru

T. Kolosova
e-mail: t.kolosova@spbu.ru

P. Eismont et al. (eds.), *Literature, Language and Computing*,
https://doi.org/10.1007/978-981-99-3604-5_14

1 Introduction

1.1 Problem Relevance

Development of syllabi in accordance with Federal State Education Standards of Higher Education (third edition) nowadays has a lot to do with practical aspects of every discipline. This is considered a most crucial challenge universities are facing.

Transition to a flexible system of learning that would take into account learners' needs and interests encourages us to look for appropriate combination of contents and formats to train future specialists.

The advancement of e-learning requires teachers to transform their professional activity, leaving room for inspiring innovations. "Transition from a traditional to an online learning format entails new functions and role of the teacher in a new educational environment. That is why revision of teacher training in Russian as a foreign language seems inevitable, taking into consideration all the finer points of online learning" [39].

Contemporary education pursues new theoretical and practical evidence-based innovations that have been expanded globally in recent years. The COVID-19 pandemic made a huge impact on the educational process, since teachers across the globe had to start actively using communication technologies in teaching and acquire new teaching models and competences. Taking the above-mentioned into consideration, it is essential to effectively build-up both general and individual professional competences of future educators [34]. As noted by M. A. Egorova, the search for answers to methodological, technological and conceptual questions posed by the current situation is of particular importance for teacher training [8]. In its turn, this implies review of the requirements to professional teacher training in the field of Russian as a foreign language.

The relevance of our research is justified by the dire need for theoretical, methodological, and organizational foundation to build professional teaching competences in Russian as a foreign language amid increasingly digitalized educational environment.

Thus, inevitably approaches to train future teachers need continuous upgrading in terms of professional standards, modern challenges, and employment requirements to ensure quality of higher education. Peer-to-peer online training is a tool to build professional competence in conformity with the current advancements in higher education.

1.2 The Goals and Objectives of the Research

The goals of the research are formulated taking into account the challenges of building professional competences among promising teachers of Russian as a foreign language in order to ensure proficient grasp of contemporary innovative education technologies, as well as additional skills required to design a vibrant linguistic environment to make it up for physical separation of the academic process participants in distance learning format. These goals are as follows:

1. to offer a detailed description of organizational set-up for peer-to-peer online teaching practice as a specific educational format in digital environment;
2. to specify individual learning as a concept;
3. to identify aspects of professional competence required today of Russian as a foreign language teachers in distance learning, to be mainstreamed within the framework of peer-to-peer online practice, as well as to specify professional skills relevant for distance learning;
4. to confirm the efficacy of peer-to-peer online practice in building the teaching qualities and competences, as well as in shaping future professional activity in digital environment.

2 Relevance

2.1 Overview of Russian Research Background

No period in the development of modern education is comparable to today's quantitative and qualitative changes [14]. Based on current standards, many researchers believe teaching implies both communication of knowledge, as well as acquisition of skills. Within the framework of systemic activity, it is essential for contemporary education to organize knowledge- and skills-intensive practice. Appropriate practice-guided technologies are especially relevant career-training of future teachers. To master the required competences undergraduates are to be engaged in cognitive activity and solve practice-based tasks [31].

G. S. Larina and A. V. Kapuza claim the main objectives in teacher training are the development of problem solving skills, ability to find one's way in the overwhelming amount of information, and make decisions in uncertain situations [19].

Theoretical and practical aspects of Russian as a foreign language studies, both in distance and off-line learning formats, are covered in research papers by E. G. Azimov, A. N. Bogomolov, A. V. Tryapel'nikov and others [3, 4, 38].

Professional teaching of Russian as a foreign language and professional skills in traditional and distance learning are reflected in the works by E. G. Azimov, A. A. Akishina, I. A. Zimnyaya, V. V. Molchanovskiy, A. N. Shchukin and others [1, 3, 24, 25, 29, 40].

Papers dealing with the influence of widespread information technologies on the job profile diagram in teaching belong to M. A. Bovtenko, E. S. Polat, S. V. Titova, etc. [5, 6, 27, 37].

Most of researchers (S. A. Deryabina, T. A. Dyakova, B. B. Mitrofanova, Z. I. Zherebtsova and others) underline the importance of IT competence in teaching Russian as a foreign language, in particular, the use of IT to improve speaking skills of undergraduates [7, 22]. A special emphasis is given to the tools (web resources, educational Internet technologies and so on) used in this work (S. S. Khromov et al., A. A. Khruleva and others) [17, 18]. These tools serve both as a didactic ground to build IT competence, and a tool for learning.

A number of contemporary researchers and instructors believe future teachers require cooperation, team work, and project management skills that appear to be incorporated into the key competences and educational standards of the twenty-first century. Another priority is building psychological and teaching competences regarding distance learning formats, including abilities "to select the most effective teaching and information conveying methods, as well as teaching materials to create a natural learning environment that would encourage productivity and the development of such individual traits as creativity, systemic analytical thinking, curiosity, self-discipline, imagination, etc." [4].

2.2 Overview of International Research Background

Laboratory and field research has proved that students working in small groups is a learning strategy that promotes socialization, cognitive and academic motivation, and learning success practically across all disciplines, starting from nursery school to college [16, 28, 30, 32].

Cooperative learning increases "students' willingness to engage in joint productive work with other people, who have various learning and adaptation needs, improving interpersonal relationship with people from different cultural and ethnic groups" [15, 33]. Co-education strategies have demonstrated ability to facilitate learning and socialization [11].

Current research argues that digital technologies allow to formalize task solving skills [13, 26].

The importance of practical activity in training future teachers is by no means underestimated in current research [35].

Digital learning should be accompanied by a systematic analysis of results based on students' feedback in order to improve postgraduate training and promote socially required professional competences [9].

Thus, analysis of methodological sources suggests it is necessary to upgrade training programs for future teachers of Russian as a foreign language, as well as to include a new practical teaching format as a professional background prerequisite.

3 Materials and Methods

3.1 Theoretical and Empirical Methods

The research methodology is based on competence activity approach, personality-oriented and communicative approach to college education organization. For theoretical generalization, we used the analysis of theoretical sources and documents, along with mass empirical data processing.

The most crucial task is to establish professional competences and relevant fundamental skills in online environment, in particular, for peer-to-peer online practice.

The mainstream method is of descriptive nature and is represented by a comprehensive set of academic approaches, allowing the researcher to discard particular empirical observations in favor of conclusions. Other methods are based on real communication and involved observation. With the consent of both teacher trainers and trainees, online lessons were recorded for further analysis and discussion. As a result, a video corpus of 20 online lessons was collected.

Questionnaire surveys are another method deployed in our research. To assess trainees' work and to receive feedback, foreign students were offered a specifically designed questionnaire.

All collected material was processed and analyzed in order to derive conceptual information. Observation ranked first as it grants get access to primary information about the collected material. Questionnaires were also analyzed using qualitative methods.

Empirical methods (practice observation and methodological analysis of online lessons) allowed to gain instant information about the competence-building process in trainees.

3.2 The Basis of the Research

Professional competence required from Russian as a foreign language teachers has undergone considerable changes. These days are no exception.

In terms of training future professionals, Russian methodology has always prioritized the development of professional skills. A person's integral linguistic mindset shall demonstrate holistic professional outlook in the relevant field, mature professional thinking and consciousness, as well as professionally relevant personal qualities.

Among general requirements to teachers of Russian as a foreign language, prominence is traditionally given to linguistic (academic knowledge of the Russian language), psychological (psychology of professional communication), methodological (methods of teaching Russian as a foreign language), pedagogical (general knowledge of pedagogy), and professional communicative competences [24, 25].

Special attention out the competences above should be paid to professional communication. Once completely established, it provides for successful communication, which is vitally important in distant learning and is primarily determined by teacher's personality, i.e. sociability, strong social and professional attitudes, creativity, dynamics, integrity, reflective ability [22].

Since teacher–student distance and online interaction is currently in the foreground of educational process, one can refer to the concept of professional personal integrity only in terms of e-learning.

The active use of electronic educational platforms allows one to confidently record changes in the job profile diagram (generalized reference model of a successful specialist) of a university teacher, in particular, to highlight new competences, including with regard to online lesson design, electronic learning materials design, and e-learning implementation.

Thus, in addition to systemic academic knowledge of the language and mastery of teaching methodology, information culture, understood as "knowledge, design of electronic courses, implementation of online types of learning" [21], becomes a vital component.

Modern innovations in the development of education system have brought up specific features into the system of postgraduate professional training. Innovations always take into account and include tradition, approaches and elements of previous methodological experience. Innovations of today are expressed primarily in the use of learning technologies, especially, e-learning technologies. Their implementation has been associated with rapid growth of the number of competencies that are necessary for effective professional activity.

It has become possible to trace the build-up of competences relevant for a modern teacher of Russian as a foreign language thanks to a new type of peer-to-peer teaching. Language teacher training programs are now actively acquiring virtual work experience in online learning. With careful guidance and deliberation, this will allow students to gain personal experience, explore various aspects of online language learning, in particular, the interaction skills.

However, virtual experience is not enough to prepare for online learning. It is necessary for students to be engaged in a virtual practical lesson. Currently existing distance learning models of Russian as a foreign language include: network training, network training and case studies, blended learning, distributed classroom learning, joint student–teacher work in simultaneous communication [39].

The latter model has given rise to a new type of practice, since it allows not only to change the process of interaction between communication participants from vertical to horizontal, transforming the roles and functions of communicants, but also to individualize the educational process.

What is meant by horizontal learning is pedagogical technology of learning in cooperation. In the system of open distance education, the idea of "horizontal" learning was embodied in the paradigm of learning activities based on the ideology of open learning resources combined with the network interaction between participants in "peer-to-peer" (or P2P) format. It is the collaborative nature of the educational process as a main characteristic of peer-to-peer approach that makes this type of learning effective.

At the same time, the teacher is a leading partner, because thanks to their activity, "the student acquires direction, intensity, forward movement and personal meaning" [10]. "It is assumed that this participant (a trainee in the role of a teacher) has the skills, abilities and information necessary to achieve a certain goal…" [2].

As for the possibility of individualizing the educational process, nowadays this topical methodological trend is being actively conceptualized and introduced into the educational paradigm. Priority of individual lessons is, first of all, determined by psychological, domestic and social factors.

The main principles of individual learning are learning flexibility and andragogy [39].

By the flexibility of training, we mean the ability to adjust the time, duration of classes, choice of teaching materials and appropriate teaching methodology, i.e. variation of the learning process depending on the goals of the student and the specific features of their psychological portrait. Flexibility is one of the most important principles of individual learning in online environment. "Learning flexibility is a transition to a system in which important decisions of the learning process are made by the students, while the teacher is ready to adapt the learning process depending on the preferences of the student, their level of communicative, computer and digital competence" [39].

The principle of training flexibility is taken into account to the fullest extent in the new type of practice. Trainee students together with their counterparts determine the learning objectives, plan the learning process (three synchronous online lessons), and make a number of important decisions (the choice of a conversational topic, its contents, their counterpart's project activity, etc.). "In a collaborative learning environment, students can criticize each other's views and opinions, as well as third-party points of view. They can turn to each other for clarifications or criticism, thus intellectually stimulating themselves and others. In addition, they can motivate and help each other in bringing the work to completion" [2].

Thus, flexibility is manifested not only in the course of the educational process, but also in teachers' (trainees') behavior and their activities. Thus, the role of teachers in educational process is gradually being transformed. They acquire new goals, tasks, methods and teaching techniques, as well as tools for interacting with students. Teaching activity is fundamentally changing. "A significant part of online education does not exclude the participation of a teacher in the educational process—on the contrary, it expands the boundaries and removes barriers in communication between a teacher and a student. "Looking for a personality" in online education should be primarily realized in the so-called simultaneous learning format, with online lessons as a major format" [20].

The second important principle of individual learning in online environment is the principle of andragogy, focused on the theory and methodology of adult learning in the context of lifelong learning. Among adult education principles, the following should be noted:

- the importance of life and professional experience;
- joint activity;
- the leading role of the student;
- practice-oriented nature of learning;
- comfortable psychological atmosphere;
- willingness to learn.

Highlighting the principles of andragogy and flexibility of the educational process as the leading ones, we establish the features of individual learning that determine the specific characteristics of the teacher's activity and, therefore, expand the job profile diagram of a Russian as a foreign language.

The effectiveness of this activity is determined by the teacher's the ability to adapt methodological approaches to educational materials and to the educational process as a whole based on the capabilities, goals and competences of the student (vis-a-vis). Such a skill is defined by the following competence:

The teacher is able to individualize learning, adapt the process of knowledge acquisition of Russian as a foreign language to a specific student, while maintaining the methodological basis characteristic of online learning.

To summarize, an individual online lesson with a teacher in the mode of simultaneous communication is characterized by:

1. an individual learning trajectory, taking into account the level of knowledge of the studied language;
2. ultimate personalization, taking into account life and professional experience of the student;
3. collaborative activity with the leading role of the student;
4. informal, friendly nature of interaction, which creates a comfortable psychological environment;
5. practice-oriented nature of learning, providing for extensive speech practice;
6. flexibility of the learning process;
7. focus on results.

All these specific features of communication and learning are aimed primarily at involving the subject of learning in cognitive activity to solve practice-oriented tasks. At the same time they demonstrate basic competences, in particular, psychological and pedagogical competences: the ability to create a natural learning environment in an online format.

The online environment of such non-formal education is systematic, flexible, competitive, customer-oriented, and efficient.

A modern teacher is increasingly viewed as a multidisciplinary specialist, expanding their positions and mastering new professions: andrologist, psychologist, educational content developer, etc. "A teacher of Russian as a foreign language should be flexible in choosing techniques, means and methods of teaching, taking into account individual and ethnic characteristics of students" [24].

Moreover, according to a number of experts, basic knowledge of teaching (linguistic and methodological) often recede into the background when it comes to specific characteristics of communication in a simultaneous online lesson with discussion as a prevailing forms of work, and give way to knowledge in psychology and communication theory.

An equally important component of professional competence within the modern digital educational space is the information literacy of the teacher, which also allows one to search for effective teaching methods and design electronic educational and methodological materials for distance learning and, therefore, ensure uninterrupted work with these online materials during the lesson.

Information literacy is understood as an information worldview, skills in the field of meeting individual information needs, as well as a set of knowledge and skills that facilitate independent design of e-courses [7].

Thus, Internet skills in education and the ability to organize teaching activities with the most productive involvement of Internet technologies is also becoming an important part of the job profile diagram for a teacher of Russian as a foreign language.

To summarize, a modern teacher of Russian as a foreign language must have a number of competences to carry out their activities in the online environment:

- group 1-

Information and communication competences, which are understood as the application of technical knowledge and skills in educational activities, the ability to correctly formulate information needs and requests, use the acquired knowledge in professional activities: organization of distance learning (designing electronic educational and methodological materials for remote interaction; organization of remote and interactive learning format in real-time mode, using effective computer methodological tools);

- group 2-

Psychological and pedagogical competencies, expressed in the ability to select effective teaching methods, ways of processing and presenting information, create a natural learning environment that contributes to productive activities and the development of the student's personality, determine the success of building communication and individualize learning, adapt the process of obtaining knowledge of Russian as a foreign language to a particular student, while maintaining the methodological basis, characteristic of online learning.

3.3 Research Basis

The manual “Methodology for conducting an online lesson within the framework of pedagogy of cooperation” presents the authors’ model, developed under the grant by Vladimir Potanin Charitable Foundation: a tripartite model of an online lesson, the concept of a constructor lesson, the form of a mind map as a summary of an online lesson. These components provide the flexibility of the educational process, taking into account students’ individual characteristics and interests, and can form the basis of the methodology of individual learning. Particular attention is paid to the difference between working with one student (individual training) and with a group.

The tripartite lesson model is based on the achievements of Russian methodology in the field of teaching Russian as a foreign language [23], as well as some provisions of innovative Russian and foreign technologies, in particular, the “Konstruktor” pedagogical technique [12, 36] adapted to the applied nature of teaching Russian as a foreign language. We believe that new forms of interaction between the teacher and the student can and should be built on the basis of a combination of traditional educational methods and the latest technologies.

The classic time format (90 min.) and the classic percentage of speech activities (speaking and listening—70%, reading—20%, writing—10%) are liberally distributed over three lessons of 30 min each.

Lesson 1. Call stage. During the lesson, existing knowledge, skills and abilities are updated.

A conversational topic relevant to the counterpart is identified and formulated (specified) on the basis of the mind map developed by the trainee. “Identification of a topic that is interesting for the student is carried out with the help of deliberate and clearly structured question-and-answer work.” The purpose of the first lesson is to identify the student’s problems in the area of grammar, vocabulary and to establish the conversational topic that is of interest to the student.

Lesson 2. Comprehension stage.

The second lesson is based on checking homework (the teacher comes into contact with the information, systematizes it, thinks about the nature of the object, and learns to formulate its position). Materials for homework are compiled by the trainee independently, taking into account student’s individual characteristics, their grammatical problems, lexical and cross-cultural gaps. Following the principles of “flipped learning”, trainees answer students’ questions, explain, correct and invite trainees to solve a number of practical problems.

Homework for the third lesson is to prepare a project that reveals the student’s point of view on the topic under discussion, in which the lexical and grammatical material updated during the second meeting is used to the maximum.

Lesson 3. Reflection stage.

New knowledge is consolidated, primary ideas are rebuilt, new concepts are included. The student “tries on” the role of a teacher. During the third, final lesson, the trainee and their counterpart change roles. The practical use of the covered material takes place in the format of a project presentation on a chosen topic, which is prepared by the student on his or her own. The work on the conversational topic is completed.

A mind map is offered as a summary of the lesson.

The radiant structure of the mind map, which imitates the characteristics of human thinking, the ability to use colors and symbols that are convenient for the trainee, the absence of a rigid structure, as well as the fact that all the necessary information is in front of the trainee's eyes at the same time, allows them to work on it without hesitation.

In the course of work within the framework of this model, students master various ways of integrating information, learn to develop their own opinions based on their understanding of versatile experience and ideas, arrive at conclusions, and confidently and accurately express their thoughts. All this becomes possible thanks to the technology of cooperation, horizontal and individual learning, accompanied by a competent use of information and communication technologies.

Competences formed in the course of peer-to-peer teaching practice are as follows:

1. ability to use e-learning materials, design learning materials for remote interaction and for online work;
2. ability to select effective teaching methods;
3. ability to design natural learning environments based on collaborative pedagogy and individual learning;
4. ability to individualize learning, to adapt the process of obtaining knowledge of Russian as a foreign language to an individual student, while maintaining the methodological basis characteristic of online learning;
5. ability to create a natural environment in online learning;
6. ability to solve unfamiliar problems, quickly navigate, make decisions in a situation of uncertainty, i.e. to anticipate, direct in a more efficient way, foresee the entire arsenal of means to achieve the goal;
7. ability to analyze and reflect. The development of pedagogical reflective thinking, critical reflective thinking regarding the pedagogical process and the development of general pedagogical culture is a necessary condition for further improvement of one's professional activity.

3.4 Research Stages

We assessed the efficiency of online teaching practice for the formation of the necessary professional competencies in future teachers of Russian as a foreign language.

A special course named "Online format of pedagogical practice", developed for first year undergraduates at the Department of Russian as a Foreign Language and Methods of its Teaching, St. Petersburg State University, acted as a methodology base. The course aims to acquaint students with modern innovative achievements in the field of Russian and foreign digital pedagogy, with innovative approaches to teaching foreign languages and, in particular, with fundamental methodology for conducting online lessons in Russian as a foreign language. In addition, it provides an opportunity to use the acquired knowledge in practice, including not only lecture materials, but also seminars on the practical projects of undergraduates. In the process of learning, students were able to master innovative teaching technologies. Implementation of a number of special tasks (drawing up a mind map of the proposed online lesson and discussing it with fellow students, conducting and discussing a fragment of an online lesson, etc.) is aimed at developing pedagogical reflective thinking, critical interpretation of the teaching process and the development of teaching culture in general.

At the first stage, the general level of students' basic theoretical and methodological knowledge on online work was assessed. Tests at the end of the course "Online form of teaching practice" showed that students developed and tested a tripartite online lesson as part of the course, followed by discussion and elaboration. Grade "Passed" was given to master's students who demonstrated confident knowledge of the special course, gave detailed and logical answers to teachers' questions related to the presented tripartite online lesson. This special course was a preparation of master's students for having peer-to-peer online practice and was aimed at developing students' professional competence.

At the second stage, students participated in peer-to-peer online practice for 30 days. The first part was receptive and included review and independent methodological analysis of 6–8 lessons on the online platform in real-time mode or recordings, followed by identification of the goals, objectives of every lesson, teaching materials and techniques, resemblance of the communicative environment to reality, and the trainees' behavior in unforeseen situations of communication.

The second part is the productive part of pedagogical practice. Each trainee was assigned a mentor who set the time for individual consultations, the purpose of which was to determine the topic and contents of lessons, as well as to elaborate the outline of the three-part online lesson developed by the trainee. Students from partner universities of St. Petersburg State University (Italy, Switzerland) acted as the recipients of the training. All the lessons were recorded for further review, discussion and methodological analysis.

Furthermore, students' work was evaluated directly during the online practice. Each trainee kept a specially designed practice diary, which also has space for the notes of the mentor. The mentor could record his or her opinion about the professional level of the undergraduate, their attitude to practice, the level of online lessons they created. Based on the results of evaluation of interns in accordance with the levels of development of the basic components of professional competence, the following figures were obtained:

Information and communication competencies:

1 application of technical knowledge and skills in educational activity—70% of trainees demonstrated an average level, 30% of trainees demonstrated a high level;

2 designing training materials for remote interaction—low level (20%), medium level (60%), high level (20%);
3 designing electronic materials for online work—low level (15%), medium level (75%), high level (10%).

Psychologists-pedagogical competencies:

4 teaching methods—medium level (15%), high level (85%);
5 ways of designing and presenting educational materials—medium level (20%), high level (80%);
6 naturalness of the learning environment—medium level (55%), high level (45%);
7 solving unfamiliar tasks, making fast decisions in situations of uncertainty—medium level (40%), high level (60%);
8 individualization of learning—medium level (40%), high level (60%);
9 analysis and reflection—medium level (20%), high level (80%).

Thus, in general, all trainees demonstrated a fairly confident level of professional competence formation, which is largely due to the coordinating actions of mentors, who ensure that the trainees' notes are checked and corrected. The methodical analysis of online lessons also contributed to identifying weaknesses and finding methodologically correct decisions in the current learning situation, which was also aimed at developing professional competence (the ability to analyze and reflect).

The obtained results allow one to state that undergraduates who participated in internship "Peer-to-peer practice at St. Petersburg State University" were able to individualize the learning process, adapt it to practical needs of the learners, create comfortable conditions of horizontal learning and to improve their Russian language skills.

4 Discussion of Results

Since the hallmark of peer-to-peer learning is focused on individual work with a student, it was necessary to clarify the concept of "individual learning". It is understood as learning according to an individual plan, drawn up taking into account the strengths of the student and their problems in mastering a foreign (Russian) language, aimed at correcting the identified problems and further improving the knowledge gained.

Despite the fact that online learning is actively used in the practice of teaching Russian as a foreign language, according to our data, there are no corresponding developments for higher education institutions. In the existing few works within this trend, the foundations of individual learning for conducting school classes have been developed. Thus, the works of Inge Unt, A. S. Granitskaya deal with methodological issues related to specific individual work with children and adolescents, its strengths and weaknesses, and provide examples from individual experience. In our work we examined the characteristics of individual work with adult learners.

We consider it important for online learning to rely on the traditional basics of Russian as a foreign language lesson. For this reason, the three-part lesson is based on the traditional time format of 90 min, and the traditional distribution of study time into four types of speech activity.

The goal of peer-to-peer pedagogical practice, which was to develop such competences in future teachers of Russian as a foreign language, that would involve mastery of modern innovative educational technologies and a special language environment design that can compensate for the separation of participants in learning communication in conditions of individual online learning, understanding its difference from traditional group classes, can be considered as achieved.

Development of the foundations of online practice and the identification and expansion of the components of professional competences of students seems to be relevant for further research. It is necessary to develop not only individual, but also group learning. Last but not least, clarification of the methodology of conducting lessons, taking into account mentors' analysis, feedback from trainees and learners should by no means be left unattended.

5 Conclusion

To summarize, a methodological base has been developed that can provide such a direction of work as organizing and conducting pedagogical online practice; the content of "professional competence" as a concept has been clarified in relation to online learning conditions, including peculiarities of pedagogical online practice by philology undergraduates of the Department of Russian as a Foreign Language and the methodology of its teaching at St. Petersburg State University in the specialty field 10.02.01 ("Russian as a foreign language") have been described. The above results confirm the formation of professional competence among students who have completed an internship "Peer-to-peer practice at St. Petersburg State University» with the following components:

- information and communication competences, which are understood as the application of technical knowledge and skills in educational activity, the ability to accurately formulate information needs and requests, use the acquired knowledge in professional activity necessary for organizing distance learning (designing electronic educational and methodological materials for remote interaction), as well as determining the possibility of implementing a remote and interactive format of learning in a real-time mode using effective computer methodological tools;
- psychological and pedagogical competences, expressed in the ability to select effective teaching methods, ways of processing and presenting information, create a natural learning environment conducive to productive activity and personal development of learners, solve unfamiliar tasks during the lesson, individualize learning, adapt the process of obtaining knowledge of Russian as a foreign language to specific practical needs of the learner, while maintaining the methodological basis, developed for online learning.

References

1. Akishina, A. A., Kagan, O. E. (2002). Uchimsya uchit. Dlya prepodavatelei russkogo yazika kak inostrannogo. [Learning to teach. For teachers of Russian as a foreign language] M., (in Russian)
2. Anciferova, O. V., Kolosova, T. N., Popova, T. I., Shchukina, K. A. (2019). Metodika provedeniya onlain-uroka v ramkah pedagogiki sotrudnichestva/Pod red. T.I. Popovoi. Uchebno-metodicheskoe posobie dlya studentov-magistrantov filologicheskih specialnostei. [The method of conducting an online lesson within the framework of the pedagogy of cooperation. Educational and methodical manual for undergraduates of philological specialties.] – SPb.: Izd-vo SPbGU, 2019. (in Russian)
3. Azimov, E. G. (2012) Informacionno-kommunikativnie tehnologii v prepodavanii russkogo yazika kak inostrannogo: metodicheskoe posobie dlya prepodavatelei russkogo yazika kak inostrannogo [Information and communication technologies in teaching Russian as a foreign language: a methodological guide for teachers of Russian as a foreign language].- M.: Russkii yazik.Kursi, 352 s. (in Russian)
4. Bogomolov, A. N. (2007) Professionalnii portret prepodavatelya v sisteme distancionnogo obucheniya [Professional portrait of a teacher in the distance learning system]. Visshee obrazovanie v Rossii, , №9, s.106–110 (in Russian)
5. Bovtenko, M. A. (2005) Kompyuternaya lingvodidaktika [Computer linguodidactics.]. M.: Flinta,. 215 s. (in Russian)
6. Bovtenko, M. A. (2005)< Informacionno-kommunikacionnie tehnologii v prepodavanii inostrannogo yazika: sozdanie elektronnih uchebnih materialov [Information and communication technologies in teaching a foreign language: creation of electronic educational materials] [Tekst] : ucheb. Posobie/M.A. Bovtenko. - Novosibirsk,. - 112s. (in Russian)
7. Deryabina, S. A., Dyakova, T. A. (2019). Professiogramma prepodavatelya inostrannogo yazika v umloviyah cifrovizacii obrazovatelnogo prostranstva [Professiogram of a foreign language teacher in the conditions of digitalization of the educational space]. Visshee obrazovanie v Rossii №4, 2019 s.142–150 (in Russian)
8. Egorova, M. A. (2017) O podgotovke kadrov v usloviyakh primeneniya professional'nykh standartov v oblasti obrazovaniya (na primere pedagoga-psikhologa). [On the training of personnel in the context of the application of professional standards in the field of education (on the example of a teacher-psychologist).] *Psikhologo-pedagogicheskie issledovaniya, 9*(3), 30–38. https://doi.org/10.17759/psyedu.2017090304 (date of access: 23.01.21). (in Russian)
9. Esquicha-Medina, A. (2018). Task-based learning in a virtual learning environment to develop writing skills in German, CEFR levels A1 and A2, in tertiary education. *Pixel-Bit, Revista de Medios y Educacion, 53*, 61–78. [Electronic resource]. Retrieved 01 May, 2022, from https://doi.org/10.12795/pixelbit.2018.i53/04
10. Fedotova, I. E. (2002). Rol prepodavatelya RKI v novih usloviyah obucheniya [The role of the Russian as a foreign language teacher in the new learning environment]. In: I. E. Fedotova (Ed.), *Inostrannie yaziki: teoriya i praktika prepodavaniya : materiali Mezhdunarodnoi nauchnoi konferencii* [*Foreign languages: theory and practice of teaching: materials of the International scientific conference*] [Minsk], 30–31 yanvarya 2002 g. / Belorusskii gos. ekon. un-t; [redkol.: T.F. Solonovich i dr.]. - Minsk : BGEU, 2002. - S. 184–186 (in Russian)
11. Gillies, R. M. (2014). Developments in cooperative learning: review of research. *Anales de psicología, 30*(3), 792–801
12. Gin, A. A. (2019). Priemi pedagogicheskoi tehniki. Svoboda vibora. Otkritost. Deyatelnost. Obratnaya svyaz. Idealnost. [Methods of pedagogical technique. Freedom of choice. Openness. Activity. Feedback. Ideality.]M.: Vita-Press (in Russian).
13. Hamada, M., Hassan, M. (2017). An interactive learning environment for information and communication theory. *Eurasia Journal of Mathematics, Science and Technology Education, 13*(1), 35–59. https://doi.org/10.12973/eurasia.2017.00603a

14. Hutorskoi, A. V. (2017). Didaktika. Uchebnik dlya vuzov. Standart tretego pokoleniya. [Didactics. Textbook for high schools. third generation standard.] – Spb.: Piter, 720 s. (in Russian)
15. Johnson, D., & Johnson, R. (2000). Cooperative learning, values, and culturally plural classrooms. In. M. Leicester, C. Modgil, & S. Modgil (Eds.), *Classroom issues: Practice, pedagogy and curriculum* (pp. 15–28). Palmer Press
16. Johnson, D., & Johnson, R. (2002). Learning together and alone: Overview and meta-analysis. *Asia Pacific Journal of Education, 22*, 95–105.
17. Khromov, S. S., Gulyaeva, N. A., Apalkov, V. G., & Nikonova, N. K. (2015). Informacionno-kommunikacionnie tehnologii v prepodavanii russkogo yazika kak inostrannogo na nachalnom etape (uroven AI, A2) [Information and communication technologies in teaching Russian as a foreign language at the initial stage (level AI, A2)] [Elektronnii resurs]. Otkritoe obrazovanie. - 2015. - № 2. - Rezhim dostupa. Retrieved 01 April, 2022, from https://cyberleninka.ru/article/ri/infonnatsionno-konmunikatsionnye-telmologii-v-prepodavanii (in Russian)
18. Khruleva, A. A. (2016) Strukturno-funkcionalnaya model formirovaniya informacionnoi kulturi budush'ih uchitelei angliiskogo yazika. Problemi sovremennogo pedagogicheskogo obrazovaniya : sbornik nauchnih statei. [Structural and functional model of the formation of information culture of future teachers of the English language. Problems of modern pedagogical education: a collection of scientific articles.] Seriya: Pedagogika i psihologiya. Vip. 52, ch. 1. - YAlta, 2016. - S. 330–339 (in Russian)
19. Larina, G. S., & Kapuza A. V. (2020) Kognitivnie processi v prepodavanii: svyaz s dostizheniyami uchash'ihsya v matematike. [Cognitive processes in teaching: connection with student achievement in mathematics.]. Voprosi obrazovaniya, 2020. № 1. S. 70–96 (in Russian)
20. Lebedeva, M. Y., & Kuvaeva, A. S. (2020) Sinhronnii onlain-urok po RKI kak osobaya forma obucheniya v cifrovoi srede. Russkii yazik za rubezhom. [Simultaneous online Russian as a foreign language lesson as a special form of learning in the digital environment. Russian language abroad.] 2020, № 2. (in Russian)
21. Mitrofanova, I. I., & Herebcova, Z. (2015). Formirovanie professionalnoi kompetencii v usloviyah cifrovizacii obrazovaniya. [Formation of professional competence in the context of digitalization of education]. Retrieved 01 April 2022, from https://cyberleninka.ru/article/n/formirovanie-professionalnoy-kompetentnosti-prepodavatelya-russkogo-yazyka-kak-inostrannogo-v-usloviyah-tsifrovizatsii-obrazovaniya? (in Russian)
22. Mitrofanova, O. D., & Kostomarov, V. G. (1988). i dr. Metodika prepodavaniya russkogo yazika kak inostrannogo. [Methods of teaching Russian as a foreign language]., M.: Russkii yazik. 270 s (in Russian)
23. Molchanovskii, V. V. (2014). Prepodavatel russkogo yazika kak inostrannogo i novie tehnologii obucheniya. Vestnik RUDN, seriya Voprosi obrazovaniya: yazik i specialnost.[Teacher of Russian as a foreign language and new teaching technologies // Bulletin of the Peoples' Friendship University of Russia, series Questions of education: language and specialty.] №1. s.19–23 (in Russian)
24. Molchanovskii, V. V. (1998). Prepodavatel russkogo yazika kak inostrannogo. Opit sistemno-funkcionalnogo analiza. [Teacher of Russian as a foreign language. Experience in system-functional analysis.] M, 320 s (in Russian)
25. Natsis, A., Papadopoulos, P. M., & Obwegeser, N. (2018). Research integration in information systems education: Students' perceptions on learning strategies, skill development, and performance. *Journal of Information Technology Education: Research., 17*, 345–363.
26. Polat, E. S. (2002). Novie pedagogicheskie i informacionnie tehnologii v sisteme obrazovaniya [New pedagogical and information technologies in the education system] [Tekst] / E.S. Polat. - M.: Akademiya. 272s. (in Russian)
27. Serrano, J., & Pons, R. (2007). Cooperative learning: We can also do it without task structure. *Intercultural Education, 18*, 215–230.
28. Shchukin, A. N. (2006). Obuchenie inostrannim yazikam: Teoriya i praktika: Uchebnoe posobie dlya prepodavatelei i studentov [Teaching foreign languages: Theory and practice: Textbook for teachers and students.] 2-e izd., ispr. i dop. - M.: Filomatis. 480 s (in Russian)

29. Sharan, Y. (2010). Cooperative learning for academic and social gains: Valued pedagogy, problematic practice. *European Journal of Education, 45*, 300–310.
30. Shmirigilova, I. B. (2018). Didakticheskaya cennost zadachi i puti ee povisheniya. Nauka i shkola. [Didactic value of the task and ways to improve it. Science and School] = № 6. – S. 130–135 (in Russian)
31. Slavin, R. (1996). Research on cooperative learning and achievement: What we know, what we need to know. *Contemporary Educational Psychology, 21*, 43–69.
32. Slavin, R., & Cooper, R. (1999). Improving intergroup relations: Lessons learned from cooperative learning programs. *Journal of Social Issues, 55*, 647–663.
33. Soboleva, E. V., Karavaev, N. L., Perevozchikova, S. M. (2017). Sovershenstvovanie soderzhaniya podgotovki uchitelei k razrabotke i primeneniyu komp'yuternykh igr v obuchenii. *Vestnik Novosibirskogo gosudar-stvennogo pedagogicheskogo universiteta.* [Improving the content of teacher training for the development and use of computer games in teaching.], *6*, 54–70 (in Russian)
34. Sullivan, P., Bobis, J., Downton, A., Livy, S., Hughes, S., Mccormick, M., & Russo, J. (2020). Ways that relentless consistency and task variation contribute to teacher and student mathematics learning, for the learning of mathematics monograph. 1.In: *Proceedings of a Symposium on Learning in Honour of Laurinda Brown* 2020 (pp 31–37)
35. Talizina, N. F. (1984). Upravlenie processom usvoeniya znanii. [Management of the process of mastering knowledge.] M.: Izd-vo Mosk. Un-ta. (in Russian)
36. Titova, C. B. (2003). Resursi i sluzhbi Internet v prepodavanii inostrannih yazikov [Internet resources and services in teaching foreign languages][Tekst]. C.B. Titova. - M.: Izd-vo MGU. - 267s. (in Russian)
37. Tryapelnikov, A. V. (2014). Integraciya informacionnih i pedagogicheskih tehnologii v obuchenii RKI (metodo- logicheskii aspekt). [Integration of information and pedagogical technologies in teaching Russian as a foreign language (methodological aspect).], 80 s (in Russian)
38. Unt I. Individualizaciya i differenciaciya obucheniya (1990). [Individualization and differentiation of training.] M., Izdatelstvo: "Pedagogika", 190 s (in Russian)
39. Zimnyaya, I. A. (2006). Obsh'aya kultura i socialno-professionalnaya kompetentnost cheloveka// Internet-zhurnal «Eidos». [General culture and socio-professional competence of a person // Internet magazine "Eidos".]. №3. www/eidos/ru/journal/2006/0504.htm (Data obrash'eniya 24.01.2018) (in Russian)
40. Zuckerman, G. A. (2020). Sovmestnoe uchebnoe dejstvie: Reshennye i nereshennye voprosy. [Co-action of Learners: Resolved and Unresolved Issues]. *Psychological Science and Education., 25*, 51–59. (in Russian).

Incorporating Informal e-Learning into Foreign Language Teaching Through Collaborative Personalization

Natalia Kucherenko, Tatiana Alexeytseva, Maria Miretina, and Olga Voicou

Abstract Modern education, more often than not, takes the path of personalization of the learning process. This personalization is better achieved through collaboration between learners and teachers, where learners take an active part in the personalization of the learning process at each stage. This paper aims to design through personalized collaboration a didactic tool that helps bring together informal and formal learning of a foreign language. Informal learning is enabled by digital technologies and originates from the daily foreign language activities already performed by university students studying French as their first or second language. These activities and the frequency of their use were revealed by the survey conducted at the first stage of the study. With the help of goal-setting and reflection cards, students were able to set their own pace, choose preferred activities and share their results. This collaborative personalization of the informal e-learning helped foster students' motivation and resulted in an increase in many informal foreign language activities especially related to production and interaction that were often overlooked by the students according to the initial survey.

Keywords Collaborative personalization · Informal learning · Formal learning · Goal-setting and reflection cards · Digital technologies

N. Kucherenko (✉) · T. Alexeytseva · M. Miretina · O. Voicou
St. Petersburg University, Universitetskaya Emb. 7/9, 199034 St. Petersburg, Russia
e-mail: n.kucherenko@spbu.ru

T. Alexeytseva
e-mail: t.aleksejtceva@spbu.ru

M. Miretina
e-mail: m.miretina@spbu.ru

O. Voicou
e-mail: o.vojku@spbu.ru

P. Eismont et al. (eds.), *Literature, Language and Computing*,
https://doi.org/10.1007/978-981-99-3604-5_15

1 Introduction

Modern education is characterized by an intensive search for the most efficient forms of educational activities and for a way to create learning and development environment that would motivate students and foster their abilities. The personalization of education plays an important part in this process.

It is common knowledge that the personalization of education occurs through educational collaboration engaging students' interests, activity and initiative, and a reflective stance of the teacher. The joint intellectual efforts by teachers and learners are aimed at mastering subject-related and soft skills and at obtaining productive outcome.

This study aims to identify ways to integrate informal learning and formal teaching to create a personalized educational environment, on the one hand, and on the other hand, create conditions for peer collaboration giving everyone the opportunity to show their individuality and uniqueness, to set and implement personally significant goals.

The novelty of this approach lies in the technology that helps personalize foreign language teaching with the help of digital technologies used by students on an everyday basis in their informal language learning, unrelated to the educational process. The access to authentic material online helps apply competences acquired in class and further motivates to learn a foreign language.

2 Theoretical Background

The theoretical framework of the study is based on two concepts: "collaborative personalization" and "informal e-learning".

A preliminary review of research requires differentiating the terms "individualized learning," "personalized learning" and "differentiated learning." Despite being near synonyms, individualized, personalized and differentiated learning are treated as separate concepts. In a differentiated or individualized classroom, the teacher takes into consideration the diverse nature of students and provides support for students who learn in different ways and at different rates and who bring to school different talents and interests. As Tomlinson mentions, teachers can accomplish differentiation by adapting the content, process and products according to their students' individual interests and learning profiles [20]. Nevertheless, the main difference between the personalized learning approach in this article and the two other concepts is the students' role in the personalization process. In individualization and differentiation, the decision making and analysis belong to the teacher, while in personalization, one of the most essential points is the responsibility of the learners. Even if the role of the teacher changes, he or she is still responsible for a work in the classroom [9].

The term 'personalized learning' can be defined differently in different educational contexts. It is used in connection with 'Personal Learning Environments' (PLEs).

PLE is a concept or an approach where technology is used to build environments of personalized learning [4]. Buckley stated that personalization can happen either by the teacher, like a natural evolution from differentiation, or by the learner. According to him, if the personalization is done by the learner, it "would require a transformation of the model of education and would change the current roles of learner and teacher." This type of personalization "would provide greater choice, responsibility and ownership in the hands of learners" [1].

The definition of personalization in education has at least five aspects. The first is "adaptation," which consists of presenting tailor-made content according to the learner's performance in a certain activity or according to his digital footprints to stay in his or her learning zone [11, 17]. The second aspect is "customization" which consists of adapting the learning content to the learner's needs, interests and experience [12, 21]. The third aspect is "individualization," where all learners follow the same learning process but at different paces [6, 7]. The fourth aspect is "differentiation" which seeks to reflect the heterogeneity of the class by focusing on "how" to deliver information to encompass different "profiles" of learners such as advanced students or students with special needs [13]. The fifth aspect is "student-centered learning" [3, 5]. Unlike the other aspects where the machine, the teacher, or the administration take charge of personalizing content, here it is the learner who takes an active part in the personalization at all levels.

Personalization can also take two forms: personalization of the process and personalization of the results [24]. The personalization of the process is the personalization of the trajectory. The end result and standards are the same for everyone, but the learning path differs from learner to learner. The differentiation is revealed at the level of the pace, the content (format and media), the final product (paper, exhibitions, examination) and the learning environment (in class, online, on an excursion, etc.). The personalization of results is a personalization of a higher level since it is a personalization without any agenda planned in advance, nor standards or common objectives, but a path, which considers strong points of the learner, their interests, aspirations and passions.

Leadbeater distinguishes "superficial" personalization from "deep" personalization. Superficial personalization provides standardized services to meet learners' needs, while deep personalization actually helps learners self-manage and organize their learning and engages them more actively in co-design and co-creation of the educational content with the help of the teacher, administrators and instructional engineers [10]. This multifaceted personalization implies, beyond its basic principles, five types of personalization: superficial, segmented, prescriptive, adaptive and collaborative personalization [22].

Superficial personalization is personalization "from the outside," at the level of representation, without any change in content or learning methods. Such as, for example, the recognition of the person's name as a kind of identity enhancement, or the personalization of a document, whether in paper or electronic form, or even at the level of the format of a text or of a picture. *Segmented* personalization consists of grouping learners into different categories according to specific standards such as region, language, "levels" (those who obtain the best or lowest results). This is a labeled personalization. *Prescriptive* personalization allows content to be presented according to the learner's profile, preferences, ambitions, aspirations and needs. This is an individualized and distinctive customization. *Adaptive* personalization is applied automatically, systematized according to a precise algorithm that makes it possible to analyze the activity of the learner in order to present him with content in response to these interactions. The more the learner interacts with the system, the more the system acquires data to manage the content in a more precise way until being able to predict activity and the behavior of the learner and his needs.

Collaborative personalization aims to ensure that the teacher and learners collaborate to develop or modify the contents of the learning. The learner is able to express his choices, to develop his own objectives. The learner selects himself/herself the technological resources to support the learning. The selection criterion may differ according to ability, curiosity and service. The learner builds an environment according to his or her experience and expertise. They can define their own network of teachers, peers and relationships to support them in their learning and guide them in their choices. Collaborative personalization implies thinking of it as transformative and non-individualistic personalization since it leads to changes resulting from collaboration and dialogue between teachers and learners. This approach supposes starting from an initiative of the learner, which can of course be prompted by a teacher or a peer, who plays an active role and makes decisions throughout the learning. Ultimately, within the collaborative system, it is the learners who influence their career, who can define their personal objectives and their resources while counting on the support of a network of people created collectively. These individual resources will not only be progressively and legitimately used by them, but they will also be at the service of the peers to enrich the shared educational content. Collaborative personalization involves learners in all stages of learning (content development, planning, resource sharing, interactions, etc.) and attributes an essential role to dialogue and human interactions. "It's this possibility of interacting with others around you, to see how the exchanges you have with others can help you build your project, give yourself an identity and move towards freedom" [15]. This form of personalization is not interested in the tool as such but in the way it is used by different partners. The environment is designed in such a way that learners can allow themselves to be more "active" and more engaged in their activities and interactions [16].

Online lifelong learning refers to the acquisition of knowledge and skills (professional and generic) seen as a continuous process, which does not end after school or university studies. It develops uninterruptedly throughout working life and continues beyond retirement, extending today to all stages of life and to all social groups, in large partly thanks to the possibilities offered by web technologies (e-learning).

Thus, lifelong learning encompasses all forms of learning:

- formal: learning activities that take place in an organized, structured and gradual context (it leads to recognized credentials such as a diploma course taken at college or university);
- non-formal: learning activities that are structured and organized in the workplace (business, government and community organizations) that are not normally

gradual and do not normally lead to a recognized credential, such as job skills and techniques for a given task required by the workplace or training but facilitate integration into the working environment;

- informal: activities that are not enough structured, self-directed and carried out at the pace of the learner, which can be linked to professional purposes such as informational content (monitoring, newsletter, etc.), useful for working or personal experience such as intergenerational learning (inside the family, neighborhood, city or community as a whole, which are also an integral part of the learning environment, just as they are part of the foundations of the economy and society). These learning opportunities must be accessible to all citizens on an ongoing basis [19].

Theory on informal and incidental learning has a central place in John Dewey's theory of learning from experience and in Kurt Lewin's field theory, wherein he developed an understanding of the way behavior changes because of the interaction of individuals and their environment. New developments in today's fast-paced workplace—globalization, uncertainty, rapid change, diversity, technology, and virtual work—raise questions about how to revise this theory. A literature review is offered to update this learning theory. Marsick and Volpe described informal learning as integrated with work and daily routines; triggered by an internal or external jolt; and an inductive process of reflection and action that is often linked to the learning of others. The acquisition of knowledge (information) was generally accomplished through self-directed learning projects. Ultimately, while these findings generally support assertions that the large majority of learning is informal and incidental, there was also a recognition of the synergy between formal and informal/incidental modes [14].

The various online devices are based on the foundations of personalized learning. It can be a structured and organized online environment that makes available to the learner and teacher (or the person supporting learning such as the tutor or the mentor) all the resources (technological and human) necessary. There are automated tools such as screening tools, learning profile identification tools, dynamic roadmaps, analysis and monitoring sheets, needs analysis questionnaires (skills or knowledge), competency frameworks (program and courses), formative (self-correcting exercises) and summative (exam and graded work) evaluation tools, generic shells for educational games, etc. [19].

Digital instruments allow more personalized forms of teaching, and forms that give the pride of place to formative activities and collaborative tasks. Therefore, we are going to keep focus on the "collaborative personalization," implemented with the help of digital technologies. Personalization is an essential element that differentiates e-learning from traditional full-time education based on face-to-face teaching, and which offers common content to be given to a group of learners without considering the specificities of each one [18, 23].

Thus, information and communication technologies play a dual role in our study: on the one hand, it is a source and means of informal learning, on the other hand, it is a way to organize and involve in personalized collaboration.

3 Materials and Research Methods

This paper uses theoretical modeling of the controlled stage of foreign language e-learning in the context of informal education and is based on collaborative personalization. The empirical study took place in the 2021/22 academic year and consisted of testing the effectiveness of integration of informal autonomous foreign language learning and formal learning using a methodological tool to set goals, to plan, to record and to evaluate and share results. Our initial assumption was that the integration of the informal foreign language e-learning and formal learning would yield better results if given a more orderly form.

At the initial stage of the study, our task was to determine what types of daily informal foreign language learning students practice using digital technologies and how often they do it. 50 1st to 4th year students of the Faculty of Philology, St. Petersburg University, took part in the survey, 30% of students having the French language as their major and 70% learning French as their second foreign language. Below are the results of the survey (Table 1).

This survey helped us identify the most popular activities, such as "listening to foreign language songs online," "watching foreign language movies/series," "following a foreign language blogger." Students could not only check frequency against each suggested extracurricular foreign language activity but also add some of their own that they practice and which are not mentioned in the table. These are some of the activities added by students: reading books online; writing essays in order to

Table 1 Types of informal language learning and frequency of their use

Type of activities			Frequency		
	Everyday	Often	Sometimes	Very rarely	Never
I listen to foreign language songs online	35	7	4	4	–
I listen to foreign language radio (audio podcasts)	5	15	15	6	9
I watch foreign language movies/series	5	26	14	3	2
I follow a foreign language blog (video, audio, text posts)	18	15	5	6	6
I watch foreign language video lessons	4	11	21	11	3
I communicate with foreign language speakers on social networks	4	6	17	7	16
I write comments in foreign languages on social networks	2	8	11	8	21
I participate in online foreign language discussion club	–	7	6	7	30
I monitor news in foreign languages using digital media	13	11	9	7	10

better apprehend language structures and to search for stylistic devices that can be used when composing texts in a foreign language; taking notes in a foreign language; translating from a foreign language and writing poetries in a foreign language; reading academic literature; using foreign language application on a laptop and/or desktop computer; listening to audiobooks; watching memes; doing foreign language searches online on subjects related to everyday life or on specific subjects; keeping personnel journal in foreign languages; watching interviews and/or nonfiction content; describing different activities and self; preparing vocabulary cards; playing digital board games and video games in foreign languages; using language learning applications; translating and abstracting everyday materials.

The survey also made it clear that certain types of activities that could occupy a bigger place in the students informal learning are not very popular, though very useful for enhancing language competences. These include "listening to foreign language radio or audio podcasts," "watching foreign language video lessons," "communicating with foreign language speakers on social networks," "writing comments in foreign languages on social networks," "participating in online foreign language discussion club," "monitoring news in foreign languages using digital media." The low popularity of these activities among students can be explained not only by the lack of interest but also by a lack of awareness about the availability of such resources or how to use them.

At the next stage of our study, we selected an experimental and a control group. The experimental group was made up of first- and third-year students of the Faculty of Philology, St. Petersburg University, who study French as their second foreign language (a total of 20 students). The didactic tool we used to organize their informal learning was "collaborative personalization." The aim of this tool is to organize and motivate autonomous and personalized daily foreign language activity. The collaborative personalization is based on the following principles:

- orientation on the student's personal interests, needs and abilities;
- a free choice of different types of activity, domains and individual pace of each student, as well as a free choice of the ways to share the results of their work;
- creativity-oriented approach, opportunity for self-realization;
- freedom associated with responsibility.

The collaborative personalization technology involves three steps. The first step includes an orientation lesson (consultation). This consultation is held at the beginning of the semester. Taking into account their existing communication experience, the teacher offers students communicative objectives (a tentative set of foreign language activities), which are classified according to the communicative language activities (perception, production, interaction and mediation) [2]. Each student chooses 3–4 objectives, enters them into a table thus setting up his or her individualized goal, specifies quantity (number of videos, articles, comments, etc.), describes the expected result and sets deadlines. Each communicative language activity contains a blank line so that the students can propose their own activity if it is not on the list. The activities should change monthly to provide variability in the daily informal learning.

Initially, two cards were used: (1) a goal-setting and planning card for informal daily learning activities; (2) and a reflection card of informal language learning. The goal-setting and planning card was supplemented with foreign language activities suggested by the students during the preliminary survey (Table 2).

The second step of the study consisted of control and discussion of interim results. Every week, each student records the performed communicative activities and selects interesting expressions and collocations, which he or she plans to use actively in the classroom; they evaluate themselves while recording the results and comment on the work performed. The reflection is carried out with the help of a reflection card (Table 3). Every student defines an individual set of foreign language activities to perform during the next month (column "My activity"). The actual performance is recorded in the second column. In the third column, students comment on the work performed, indicate the sources, and in the fourth column, they enter useful, in their opinion, language material, which they plan to actively use. Once a month, information is shared in class. The students share the most interesting sources of information and useful activities [8].

In order to engage students in a collaborative process and share information more quickly and efficiently, we have created a shared google table for the goal-setting

Table 2 Goal-setting and planning card for informal daily learning activities

Type of activities	My choice	Deadline/frequency
I listen to foreign language songs online		
I listen to foreign language radio (audio podcasts)		
I watch foreign language movies/series		
I monitor news in foreign languages using digital media		
I read foreign language books online		
I write comments in foreign languages on social networks		
I communicate with foreign language speakers on social networks		
I participate in online foreign language discussion club		
I use language learning applications		
I play digital board games and video games in foreign languages		
I translate and abstract documents from everyday life		

Table 3 Reflection card of informal language learning

My activity	Self-assessment	Comments	Language material
Week 1			
Week 2			
Week 3			
Week 4			

card and the reflection card. Each student has a personal page and at the beginning of each month, all group members enter their chosen activities, frequency and deadlines. Self-assessment, results, and comments are added as the activities are carried out.

At the final step (at the end of the semester), students present the results of their activities, they discuss the advantages and disadvantages of various information resources, applications and platforms and sum up the results.

4 Findings

In addition to the experimental group and to compare the results, we interviewed the control group of students who did not work with collaborative personalization tool and did not use the reflection cards. The control group was composed of 10 s-year students of the same faculty with French as their major. It is assumed that initially the students of the control group were more motivated. This can be explained by the greater number of hours devoted to learning French and, consequently, by a higher level of communicative competences. As our study focuses on the progress made, compared to the primary indicators, we consider a slight difference in the initial parameters between the experimental and control groups acceptable.

We divided the obtained results into two categories. The first category included the most frequently used informal learning activities (Table 4), which previously, before the experiment, were present in the everyday life of most students. The figures show that for most of the foreign language activities there has been an increase in their usage. A significant increase (+20%) was recorded in such activities as watching movies in a foreign language; reading and watching the posts of bloggers who are native speakers of a foreign language; watching videos devoted to foreign language learning and using mobile applications for foreign language learning. No or little increase has been observed for such activities as listening to radio (audio podcasts) in a foreign language, monitoring news in foreign-language online media, and reading books online.

We put the least frequently used foreign language activities into the second category (Table 5). They were practiced by a minority of students or were not present at all in their everyday life. A significant increase in the figures is explained by the fact that students had to regularly choose new activities. A motivating factor for them was the exchange in the group with other students, which allowed them to learn about new information resources, as well as previously unknown to them ways of communicating with native speakers.

Summing up all the data presented above, we can state that students use a significant number of receptive types of foreign-language communication in their everyday informal learning activities. As it becomes also evident, there is a clear interest in online interaction. Mediation and productive activities are practiced to a lesser extent, which is apparently due to the less developed communicative skills of students learning French as their second language. These findings will allow us in the future to better focus on motivating students to use more often productive, mediative, and

Table 4 The most frequently used informal language e-learning activities

Type of activities	Before the experiment %	After the experiment %	Control group %
I listen to foreign language songs online	100	100	100
I listen to foreign language radio (audio podcasts)	60	60	70
I watch foreign language movies/series	70	90	90
I follow a foreign language blog (video, audio, text posts)	80	100	90
I watch foreign language video lessons	70	90	80
I monitor news in foreign languages using digital media	60	70	60
I read foreign language books online	70	70	70
I use language learning applications	60	80	80

Table 5 The least frequently used informal e-learning activities

Type of activities	Before the experiment %	After the experiment %	Control group %
I communicate with foreign language speakers on social networks	50	70	50
I write comments in foreign languages on social networks	40	60	50
I participate in online foreign language discussion club	20	60	30
I play digital board games and video games in foreign languages	30	70	10
I translate and abstract documents from everyday life	30	50	40

interactive forms of communication in their informal learning. For example, we can introduce a rule that among the activities chosen by students, there should be at least one production and/or mediation activity.

5 Conclusion

The results of the study indicate that informal learning with the use of digital technologies has significant potential for the personalization of foreign language formal learning. Although the study was conducted on the basis of French, the recommendations can be applied when teaching other languages.

The use of collaborative personalization can be a promising way to increase motivation to learn foreign languages. It allows personalization of the process of language learning by transferring acquired language skills and/or competences to an extracurricular context. A higher motivation is achieved due to the fact that students choose for themselves the activities they are interested in. The deadlines, the need to fill out a reflection card and the necessity to share information open up a variety of opportunities for integrating individual informal learning and formal classroom activities.

The data obtained showed that the use of collaborative personalization and reflection cards provided an increase in most types of foreign language activities. Thanks to the exchange of information about different resources and sharing of opinions on the efficiency of different types of communicative activities, students started to show more interest in the activities, which before the experiment were overlooked or completely avoided by some students.

References

1. Buckley, D. (2006). *The personalisation by pieces framework: A framework for the incremental transformation of pedagogy towards greater learner empowerment in schools.* CEA Publishing.
2. Council of Europe. (2017). *Common European framework of reference for languages: Learning, teaching, assessment*. Council of Europe Publishing, Strasbourg.
3. Demski, J. (2012). This time it's personal – T.H.E. *Journal, XXXIX*(1), 32–36
4. Fiedler, S., Väljataga, T. (2011). Personal learning environments: Concept or technology? *International Journal of Virtual and Personal Learning Environments, 2*(4), 1–11. http://ejournals.ebsco.com/Article.asp?ContributionID=27745285.
5. Kallick, B., Zmuda, A. (2017). Students at the center: personalized learning with habits of mind. Alexandria, Virginia: ASCD
6. Karpinska-Musial, B., Dziedziczak-Foltyn, A. (2014). At students' service. Tutoring and coaching as innovative methods of academic education in Poland. In: *6th International Conference on Education and New Learning Technologies (EDULEARN)* (pp. 6057–6064). Barcelona, Edulearn
7. Keefe, J. W. (2007). What is personalization?» – phi delta Kappan*, LXXXIX*(3), 217–223
8. Kolesnikov, A. A., Liskina, T. V. (2019) Guided self-study of foreign languages in the context of informal education: theoretical foundations and empirical research among students of the Institute of Foreign Languages. *Language and Culture* 48. https://doi.org/10.17223/19996195/48/18
9. Kuutila, N. (2016) Personalised learning in English as a foreign language education. Retrieved 05 September, 2022, from https://jyx.jyu.fi/bitstream/handle/123456789/50973/URN:NBN:fi:jyu-201608193829.pdf
10. Leadbeater, C. (2006). The future of public services: Personalised learning. In: *Personalising education* (pp. 101–114). Paris, OECD
11. Lefèvre, M., Guin, N., Jean-Daubias, S. (2012). Personnaliser des activités pédagogiques de manière unifiée: Une solution à la diversité des dispositifs, STICEF. Retrieved 08 April, 2022, from http://sticef.univ-lemans.fr/num/vol2012/09-lefevre-individualisation/sticef_2012_NS_lefevre_09p.pdf
12. Lin, C. F., Yeh, Y.-C., Hung, Y. H., Chang, R. I. (2013). Data mining for providing a personalized learning path in creativity: An application of decision trees. *Computers & Education, LXVIII*, 199–210
13. Looi, C.-K., et al. (2009). Anatomy of a mobilized lesson: Learning my way. *Computers & Education LIII, 4*, 1120–1132. https://doi.org/10.1016/j.compedu.2009.05.021

14. Marsick, V. J., Watkins, K. E., Callahan, M. W., Volpe, M. (2006). Reviewing theory and research on informal and incidental learning. Retrieved 08 April, 2022, from https://files.eric.ed.gov/fulltext/ED492754.pdf
15. Meirieu, P. (2011). Innover dans l'école: pourquoi? Comment? Retrieved 10 April, 2022, from http://www.meirieu.com/ARTICLES/innoverdanslecole.pdf
16. Meirieu, P. (2011). Autonomie et Apprentissage. Retrieved 10 April, 2022, from http://gfph.dpi-europe.org/comptesrendus/compterenduspairemulation/GFPH2012/autonomie-et-apprentissage.
17. Merzeau, L. (2012). La médiation identitaire. Revue Française des Sciences de l'Information et de la Communication 1. Retrieved 10 April, 2022, from http://rfsic.revues.org/193
18. Nabizadeh, A. H., Leal, J. P., Rafsanjani, H. N., Shah, R. R. (2020). Learning path personalization and recommendation methods: A survey of the state-of-the-art. *Expert Systems with Applications, 159*. https://doi.org/10.1016/j.eswa.2020.113596
19. Sauvé, L. (2014). Des dispositifs en ligne pour personnaliser l'apprentissage tout au long de la vie: Quelques recommandations. *Distances et médiations des savoirs*. https://doi.org/10.4000/dms.629
20. Tomlinson, C. A. (2014). The differentiated classroom: Responding to the needs of all learners. Alexandria, VA: ASCD
21. Watson, W. R., Watson, S. L., Reigeluth, C. M. (2013). Education 3.0: breaking the mold with technology. *Interactive Learning Environments*. https://doi.org/10.1080/10494820.2013.764322
22. Youssef, E., & Audran, J. (2019). La personnalisation de l'apprentissage vue comme facteur effectif d'innovation pédagogique. *Spirale - Revue de recherches en éducation, 63*(1), 157. https://doi.org/10.3917/spir.063.0157
23. Zhang, L., Basham, J. D., & Yang, S. (2020). Understanding the implementation of personalized learning: A research synthesis. *Educational Research Review, 31*. https://doi.org/10.1016/j.edurev.2020.100339
24. Zhao, Y. (2015). Personalization and autonomy. In: Y. Zhao, H. Tavangar, E. McCarren, G. F. Rshaid and K. Tucker (Eds.), *The take-action guide to world class learners. How to make personalization and student autonomy happen* (pp. 8–18). Corwin Press, California

Dental Internet Resources in Teaching Specialty Language Within Russian as a Foreign Language Course

Anna Yu. Tiraspolskaya

Abstract The article is devoted to educational application of dental Internet resources in teaching the specialty language within Russian as a foreign language course. At the present stage formation of students' professional communicative competence in the teaching of Russian as a foreign language is impossible without the use of information and communication technologies. The article describes options to incorporate information and educational Russian websites, forums, and chats for dentists into teaching dental terminology to foreign students. Analysis of scientific and methodological literature reveals that teaching Russian as a foreign language, including specialty language, to prospective dentists is underdeveloped. In this article, we focus on professional communication between doctors and future doctors (teacher – student format), or between international students, or colleagues in formal, semiformal, or informal setting to discuss academic, educational, scientific, and professional issues. Equally relevant is the use of professional Internet resources as an extra tool to teach Russian as a foreign language. Active involvement of foreign students in professional communication at forums and chats where dentists discuss practical and scientific problems, as well as the tasks offered by their teachers allows to significantly improve students' communicative competence in specialty language.

Keywords Russian as a foreign language · Specialty language · Dentistry · Internet resources · Communicative competence · Professional communication

1 Introduction

The life of modern man is unthinkable without the use of rapidly developing digital technologies: they have a powerful influence on all areas of society, including education. As a *force majeure*, the pandemic compelled national educational institutions—primarily universities—to scale up distance learning in the shortest possible time, meanwhile opening up new ways for the development of higher education

A. Yu. Tiraspolskaya (✉)
St. Petersburg University, Universitetskaya Emb. 7/9, 199034 St. Petersburg, Russia
e-mail: a.tiraspolskaya@spbu.ru

P. Eismont et al. (eds.), *Literature, Language and Computing*,
https://doi.org/10.1007/978-981-99-3604-5_16

[6, 26]. In particular, this forced teachers to pay close attention to various Internet resources, both educational, professional, and even semi-professional, when looking for publicly available materials and cutting-edge learning formats for different activities and assignments, for example, specific assignments aimed at maintaining students' cognitive interest (for more information see: [29: 113]).

When teaching Russian as a foreign language in new learning conditions, it is impossible to ignore one of the most serious challenges, i.e. shaping students' communicative competence in the framework of professional communication. Experience shows that acquisition of this competence—which is deemed to demonstrate the highest level of language proficiency [1: 6]—turns out to especially suffers within the distant learning format. In this situation teachers and students are assisted by Internet resources on specific subject matter of study. This article discusses the educational potential of professional dental websites to teach Russian as a foreign language to future dentists at an advanced stage. We suggest to focus on ways to shape educational, scientific, and professional communicative competences, highly demanded by senior university students.

2 Literature Review

A surge of interest among domestic researchers and methodologists to distant learning (including in higher education) with the active use of information technologies has been observed since the early 2000s. In particular, it is necessary to recall the monograph by V.P. Demkin and G.V. Mozhayeva "Tekhnologii distantsionnogo obucheniya" ["Distance Learning Technologies"] [7], which comprehensively considered the problems of organizing educational process in distant learning format, studied didactic models, analyzed the efficiency of pedagogical technologies, and other works (see: [21]). During the same period, studies are published to explore distant learning in foreign languages [18, 30]. Since the second half of the 2000s, Russian science has seen a veritable abundance of works analyzing, developing, and enhancing information and communication technologies in distant or blended learning formats [8, 14–16, 22, 23] (note that in these years the concept "distance learning" itself has been clarified and verified, fore information see e.g. [20]), as well as in "traditional" classroom-based face-to-face learning with IT assistance [4, 12].

The last decade has seen a breakthrough in the development of information technologies in all spheres, including education, thus producing numerous scientific and methodological works on how to implement some of the latest Internet services in teaching (in particular, for foreign languages, including Russian as a foreign language), with a special focus on interactive two-way communication with users. These works can be split into several groups by most characteristic examples. The first group includes articles where authors analyze Internet educational resources (programs, applications, and websites) in Russian as a foreign language, indicating their benefits and disadvantages [31]; the second group includes studies devoted to the use of video conference format in the learning process [28]. The third group describes how to apply language and other podcasts at foreign language classes [27], including Russian as a foreign language [24]; the fourth group focuses on web forums and their application at language lessons [19]; those include chats, chatbots [3], Internet messengers [9, 17], as well as intelligent voice assistants [2].

At the same time, if we turn to the analysis of scientific and methodological literature on versatile application of informatization tools in teaching Russian professional medical language to foreign students specializing in medicine and dentistry, we will see that material is extremely scarce. Strictly speaking, there are few works devoted to teaching Russian as a foreign language with a focus on specialized medical language or, more so, the language of dentistry. Most available body of research builds on classical methods of teaching Russian as a foreign language, such as selecting vocabulary or texts related to professional topics, learning terminology, mastering grammatical structures most relevant to the area of professional communication [13], active and interactive learning formats [10, 25] creating textbooks for dedicated training [5], etc. All that taken together fails to address the problems related to application of information technologies. A representative example is the article by Goncharenko “RKI dlya studentov-medikov: aktivnye i interaktivnye formy obucheniya” [“Russian as a foreign language for medical students: active and interactive forms of learning”] [10], that presents the communicative-interactive approach (for diverse formats mentioned above) without taking information and communication technologies into account. The work by Itinson and Grishkun [11] deserves special attention. Currently this work is almost an only comprehensive study to disclose the potential informatization tools in teaching medical Russian language to foreign students of medical faculties at all stages of training.

3 Materials and Methods

As the review of scientific and methodological works shows, currently application of information and communication technologies in teaching Russian dentistry language to foreign students majoring in dentistry remains undeveloped for all formats—full-time, mixed, and distance learning. On the other hand, research on methods for foreign students to master professional communication in dentistry cannot be called exhaustive. For example, most methodologies in teaching Russian as a foreign language rely on “doctor-patient” or “doctor-nurse” communication as a routine professional setting. However, we must not forget that learning a major equally entails professional communication between the student and the educator who is a professional dentist or among students. Thus, within the first five years of professional studies future dentists are challenged by formal, semi-formal or informal professional communication with colleagues in educational professional, or scientific setting. This means that the objective for a Russian as a foreign language teacher is to comprehensively train linguistic, sociolinguistic, strategi, discursive, as well as dedicated professional communicative competencies. The ability to ask a teacher a question about the content of a specialized lecture, to participate in a discussion among professionals, to agree or disagree with opposing views, or to justify one's position regarding a scientific or practical issue—dental students shall acquire these skills if he/she wants to become a true professional in their field when doing studies in Russian. At the same time, without a dedicated dental education, teachers of Russian as a foreign language may experience difficulties from time to time due to lack of subject knowledge and—more so—professional competence, that there are delegated to promote in their students. Here teachers and students can refute to strategies, that are not entirely appreciated by modern teaching methods—i.e., obtain information, training, or consultations with professional dental Internet sites in Russian.

It should be emphasized that this article does not aim to describe specific dental websites. We refer to a few most popular professional resources to highlight only some "typical" features and Internet options that may be useful for foreign dental students, learning Russian.

The purpose of the study is to describe and analyze their application in learning the language of dentistry within Russian as a foreign language course. Such websites can provide students and teachers with options for self-education, as well as an auxiliary teaching tool. Particular attention is offered to the role of dental forums and chats for learning purposes, as we describe types of exercises compiled using such Internet resources to be offered by teachers to foreign dental students in order to improve oral and written communicative competence in the field of dentistry.

4 Results

Russian dental Internet sites and forums, that are popular among professionals and students, regardless of their general content and educational or advisory potential, provide ample opportunities for dental students to expand their knowledge of professional Russian language and to improve professional communication skills and abilities under the guidance of a Russian as a foreign language teacher.

In higher education, international students (especially in their first years) are not always able to realize their professional communication needs for a variety of reasons, ranging from communication barriers and fear of communicative failure on the lack of time to communicate teachers and peer students in some cases. The communication in Internet forums or chats allows them to relieve psychological tension by, firstly, creating familiar communicative environments and secondly, offering anonymous or pseudonym-disguised conversations.

4.1 Opportunities to Use Professional Dental Forums

As a (predominantly) asynchronous means of communication, professional forums allow students to think through and formulate their questions and answers without requiring immediate response. That is why participation in dental forums should be recommended as an extra tool to teach professional communication to foreign students already in their first years. Given that most of the forums contain sections aimed at different types of audience and imply varying degrees of sophistication in presenting dental problems (i.e. "For Physicians", "For Residents/Students", and "For Patients"), students with different levels of language ability can participate in such conversations.

To get the utmost benefit from professional forums, it seems appropriate that the choice of relevant Internet resources should be made by the teacher, who would evaluate them according to the following criteria in advance: linguistic correctness (literacy) of professional language, clarity, logical correctness of statements, comprehensive comments, attitudes to interlocutors' opinions, complexity of statements (general knowledge), complexity of statements in terms of professional communication.

A successful example is the authoritative "All-Russian Dental Forum" (https://stom.ru/) [Vserossijskij stomatologicheskij forum], which is very popular among dentists and patients and undoubtedly can be recommended as a source of material for Russian as a foreign language classes for international dentists of different professional training level. This forum has a clear distribution of conversations by type of communicator and purpose of communication: there is a block of conversations called "For patients", where non-professionals can get expert advice regarding dental problems of interest; there is a block called "For doctors", designed to discuss issues between professionals; and there is also a separate block called "For students", where prospective dentists can consult on practical and academic issues. "For Patients" and "For Doctors" blocks show discussions divided into narrowly specialized areas of dentistry: "Therapy", "Orthopedics", "Dental Surgery", "Orthodontics", "Pediatric Dentistry"; there is a "Forum for Dental Technicians" and a "General Questions" section. The format of communication provides for detailed questions and answers; the expertise level is high (according to reviews by dentists), and communication is performed in fairly correct Russian language, using comprehensively structured statements.

Another forum that is respected in the professional community is the "Dental Forum (Dental Forum)" (https://stomatologclub.ru/forum/) [Forum stomatologov (stomatologicheskij forum)] at the "Dental Club" website. In contrast to the "All-Russian Dental Forum" this forum does not have separate thematic blocks for patients and professionals, but its discussions are also split across different areas of dentistry, including periodontics and implantology. There is a constant flow of live communication between the participants on the forum platform (for example, a discussion in a popular subsection can produce 500 to 2,000 posts!). The range of topics discussed is very wide, including active disputes between professionals showing especially valuable content for international students.

Classes of Russian as a foreign language offer students opportunities to familiarize with the most interesting questions of forums, to analyze opinions of different experts from the point of view of professional correctness, to compare answers, to state their point of view on a dental problem, to perform teamwork on lexical and grammatical aspects of specialty language.

Teaching communicative culture of professional communication in written form can also benefit from forum materials. In particular, the web site section “Ask expert” is very popular with students because by asking relevant questions to dental professionals, foreigners can both improve their subject-professional competence in different specialty areas, but also practice most necessary skills of professional oral and written speech in different situations, such as a detailed description of complaints (feelings of the patient), disease symptoms and X-ray results, formulation of diagnosis, etc. For specialty language classes, such forum sections and written communication between students and professionals allow to successfully practice speech patterns for certain professional situations. Students talking to experienced dentists on a medical case within the forum setting are the basis for elaborate student dialogues and polylogues, allowing to organize mini-discussions and conversations on professional topics. Independent assignments may require students to prepare brief written educational or scholarly reviews using various forum sections (e.g., “Stomatitis,” “Wisdom Tooth Removal,” “Deep Caries,” “Gum Festering,” “Gum Bleeding”) and its thematic “branches” (individual discussions of particular issues), followed by an oral presentation of the review in class.

Forums sections with videos in the field of dentistry are, perhaps, most valuable for to develop oral and written professional speech skills in future dentists. As a rule, such videos may or may not include audio commentary and demonstrate classic, difficult or interesting clinical cases encountered by a highly qualified specialist (dental treatment, extraction, prosthetics, surgery, etc.). Sometimes the author of a video gives a short written commentary below the video. In addition to the opportunity to address specialists in writing and ask questions about the content of the video or to practice listening and understanding of authentic texts in the specialty field, students have other options at hand. For example, students having advanced specialty language proficiency may carefully watch videos several times in mute mode and prepare a detailed written description of dental work, intensively using dental language, and then comment verbally on all the steps of the demonstrated procedure when watching the video another time.

4.2 Opportunities to Use Professionally-Oriented Internet Chats

As a synchronous means of communication, chats can be used in specialty language learning as an effective auxiliary tool to master advanced written and oral modes of informal communication in a professional environment. As is known, one of the primary tasks for a teacher of Russian as a foreign language is to create communicative situations in the classroom that would resemble natural communicative setting: in this way only can we build-up and improve students’ communicative competence. Specialized Internet chats are especially valuable in studying specialized language because they engage students in a lively, natural process of professional communication, instead of mimicking it. Chats tends to create a more comfortable (compared to classrooms) psychological atmosphere, which helps students combat their linguistic insecurities.

The content of chats can be divided into general, dedicated to all problems of dentistry, and highly specialized (therapeutic chats, orthodontic chats, chats for dental surgeons, chats for dental technicians, etc.).

For example, it is worth highlighting the largest and most popular among Russian dentists Telegram-chat “Dentistry” of general scope (https://telegram.me/stompub) [Stomatologiya], where students can promptly get brief answers to questions on dentistry, express their opinions and share their experiences. This chat is intended for specialists only, so the communication in it is characterized by a high-level professional language, full of terminology. Such communication, though challenging for foreigners, promotes intensive acquisition of conversational dentistry language. This system also has highly specialized chats: therapeutic (https://t.me/stomter), chats for dental surgeons and implantologists (https://t.me/stomsur), orthodontic (https://t.me/stomort), chats for dental technicians (https://t.me/zubteh) and even chats for students with more accessible content and language (https://t.me/stomstud).

The second type of chats (highly specialized) seems most useful for residency or graduate students, while the first type (general ones) can be used effectively by 1st to 5th year students. Student chats are a category of their own, offering particular activities on exchanging educational and scholarly information, arguments, and professional conversations. Chats designed for the students are assumably the most popular among foreigners thanks to relaxed and psychologically comfortable communication with communicators of equal status and less specialized level of knowledge allowing foreigners to overcome language barriers as successfully as possible. As early as their junior years, dental students can develop and consolidate communicative skills in specialized language, as well as ability to practice challenging grammatical structures and vocabulary in natural communication in such chats. As additional assignments, students can prepare written or oral reports, i.e., retell the content of a discussion or describe different points of view expressed by communicators on an issue, or prepare small reports on their chat communication on a particular professional topic.

It should be noted that now an increasing number of chats also include the option to send voice messages to the recipients. This option can undoubtedly help foreign students gain additional listening and especially speaking skills in informal professional communication, which modern students (including in the format of distance learning) often lack in their university training.

5 Discussions

According to the author’s observations and results of our student survey, communication in professional dental forums and chats in conjunction with the types of tasks described above during one academic year contributed to: development of students’ professional erudition; elimination of existing communicative barriers and mastering culture of debate on professional issues; thus significantly replenishing dental terminology lexicon, developing automatic use of basic grammatical constructions in professional contexts. At the same time, it seems that the listed results could have been even higher if the educational work described in the article had been carried out with students on the basis of forums and chats of Internet resources controlled by the university or the system of medical and dental universities or faculties, and, in particular, if they imposed more standardized requirements to the linguistic correctness of the participants (communicators) statements. So, for example, within the forum it is possible to ban remarks of participants consisting of less than 5 significant words, to equip the site with a complex text editor, including professional dental vocabulary (terminology), and so on.

Establishing of a specific Internet resource on the university basis to promote live professional communication and discussion of educational and scientific issues in dentistry with the options of communication in the following formats: 1. "lecturer-student" (where students can ask questions to the lecturer/expert) 2. "student-student" (where students can freely exchange knowledge, discuss theoretical and practical educational dentistry problems among themselves) would greatly facilitate teaching Russian as a foreign language. The availability of such a resource, including forums and chats, would allow teachers to take tighter and more consistent grasp over informal professional communication of foreign students mastering Russian, namely: to monitor linguistic aspects of professional communication of students outside the classroom more quickly and systematically; to control accuracy of their speech; based on the content of the forums, create an advanced and consistent system of assignments to improve specialty language (lexical, grammatical and syntactical aspects) that would both involve students' participation in discussions on specific topics, but would also precede these discussions, which would make it much easier for foreigners to communicate with students and professionals, both in writing and orally.

6 Conclusions

The Searching and selecting professionally oriented sites, forums, and chats would help teachers enrich and diversify the content of classes for basic Russian as a foreign language taught to dentists, which is a uneasy, time-consuming process crucial to prepare students for further working environments. However, thoughtful selection of Internet resources in accordance with the most important specialty language learning criteria provides students with broad prospects to practice communicative skills and abilities, successfully stimulate their cognitive interest both to the contents of professional learning and to relevant communicative opportunities to deeply and comprehensively master the language of dental science.

The next most important teaching objective is to create blocks of oral and written assignments to accompany students throughout their work with dental Internet resources, from “passive” acquaintance with the content of websites, forums and chat rooms to active participation in written and oral professional communication in the dental environment. A system of simulation assignments helps students to acquire skills in informal professional on-line communication, as well as to learn how to extract linguistic information themselves from dental resources (replenishing the “specialized vocabulary” with new terms, identifying grammatical structures relevant to the specialty language, identifying appropriate situations to use them in speech, etc.). These competencies help students to actively learn the language of dentistry throughout their professional life in a Russian-speaking environment.

University Internet site, forums and chats give extra focus on academic and scientific communication among dental students both with their peers and with specialists within one or more universities. Professional conceptual and linguistic supervision would greatly increase the efficiency of such Internet resources in teaching Russian as a foreign language to international dental students.

Sources

1. Forum stomatologov (stomatologicheskij forum) [Dental Forum (Dental Forum)], https://stomatologclub.ru/forum/, last accessed 2022/07/31.
2. Stomatologiya [Dentistry], https://telegram.me/stompub, last accessed 2022/08/19.
3. Vserossijskij stomatologicheskij forum [All-Russian Dental Forum], https://stom.ru/, last accessed 2022/08/25.

References

1. Akishina, A. A., & Kagan, O. E. (2002). *Uchimsya uchit’: Dlya prepodavatelya russkogo yazyka kak inostrannogo* [Learning to teach: For the teacher of Russian as a foreign language] (In Russian) (2nd ed.) revised and extended. Russkij yazyk. Kursy.
2. Al-Kaysi, A. N., Arkhangelskaya, L. V., & Rudenko-Morgun, O. I. (2019). Intellektual’nyj golosovoj pomoshhnik Alisa na urokakh russkogo yazyka kak inostrannogo (Uroven’ A1) [The intellectual voice assistant Alice in the lessons of Russian as a foreign language (Level A1)] (In Russian). *Filologicheskie nauki. Voprosy teorii i praktiki* [Philological Sciences. Questions of Theory and Practice]*, 12*(2), 239–244.
3. Bikkulova, O. S., & Ivkina, M. I. (2021). Chat-bot v metodike prepodavaniya RKI [Chat-bot in the methodology of teaching Russian as a foreign language] (In Russian). *Mir russkogo slova* [The World of Russian Word]*, 1*, 91–96.
4. Chernukhina, N. V. (2014). Informatsionno-kommunikatsionnye tekhnologii v obrazovatel’nom protsesse vuza [Information and communication technologies in the educational process of the university] (In Russian). *Nauchno-metodicheskij ehlektronnyj zhurnal «Kontsept»* [Scientific and Methodical Electronic Journal “Concept”]*, 30*, 51–55. Retrieved February 07, 2022, from http://e-koncept.ru/2014/14861
5. Chirkova, V. M. (2019). Obuchenie yazyku spetsial’nosti inostrannykh studentov-stomatologov [Teaching the language of specialty to foreign dental students] (In Russian). *Karel’skij nauchnyj zhurnal* [Karelian Scientific Journal]*, 8*(2) (27), 74–76.

6. Danilova, L. N. (2020). Covid-19 kak faktor razvitiya obrazovaniya: perspektivy tsifrovizatsii i distantsionnogo obucheniya [Covid-19 as a factor in the development of education: prospects for digitalization and distance learning] (In Russian). *Vestnik Surgutskogo gosudarstvennogo pedagogicheskogo universiteta* [Bulletin of Surgut State Pedagogical University], *5* (68), 124–135.
7. Demkin, V. P., & Mozhaeva, G. V. (2003). *Tekhnologii distantsionnogo obucheniya* [Technologies of distance learning] (In Russian). Izd-vo Tomskogo un-ta, Tomsk.
8. Devterova, Z. R. (2011). Organizatsionnye formy distantsionnogo obucheniya i spetsifika ikh primeneniya v informatsionno-obrazovatel'noj srede [Organizational forms of distance learning and specifics of their application in the information and education environment] (In Russian). *Sibirskij pedagogicheskij zhurnal* [Siberian Pedagogical Journal], *12*, 79–87. Novosibirskij gosudarstvennyj pedagogicheskij universitet, Novosibirsk.
9. Gatulin, R. R., & Kolupaeva, D. A. (2017). Ispol'zovanie messendzhera Telegram dlya realizatsii tekhnologii ehlektronnogo obucheniya v vuze [Using Telegram messenger to implement e-learning technology in higher education] (In Russian). *Sankt-Peterburgskij obrazovatel'nyj vestnik* [St. Petersburg Educational Bulletin], (11–12): 15–16, 31–33.
10. Goncharenko, N. V. (2014). RKI dlya studentov-medikov: aktivnye i interaktivnye formy obucheniya [Russian as a foreign language for medical students: active and interactive forms of teaching] (In Russian). *Russkij yazyk za rubezhom* [Russian Language Abroad], *5*, 4–6.
11. Itinson, K. S., & Grishkun, V. B. (2019). Vozmozhnosti i preimushhestva vzaimosvyazannogo primeneniya sredstv informatizatsii pri podgotovke po russkomu yazyku inostrannykh studentov v meditsinskom vuze [Opportunities and Benefits of Interconnected Application of Informatization Tools in Russian Language Training for Foreign Students in Medical Universities] (In Russian). *Vestnik RUDN. Seriya: Informatizatsiya obrazovaniya* [RUDN Journal of Informatization in Education], *16*(2), 103–116.
12. Konyukhov, M. I., & Zykin, S. A. (2016). Ispol'zovanie informatsionnykh tekhnologij distantsionnogo obucheniya pri ochnom obuchenii [The use of information technologies for distance learning in full-time education] (In Russian). *Mezhdunarodnyj zhurnal gumanitarnykh i estestvennykh nauk* [International Journal of Humanities and Natural Sciences], *1*(8), 98–100.
13. Korobkova, A. V. (2009). Formirovanie grammaticheskikh navykov u inostrannykh studentov-stomatologov [Formation of grammar skills of foreign dental students] (In Russian). *Sibirskij pedagogicheskij zhurnal* [Siberian Pedagogical Journal], *3*, 121–126.
14. Krylova, E. A. (2020). Tekhnologiya smeshannogo obucheniya v sisteme vysshego obrazovaniya [Technology of blended learning in the system of higher education] (In Russian). *Vestnik tomskogo gosudarstvennogo pedagogicheskogo universiteta.* (TSPU Bulletin) [Bulletin of Tomsk State Pedagogical University. (TSPU Bulletin)], *1* (207), 86–93.
15. Malinina, I. A. (2013). Primenenie tekhnologij smeshannogo obucheniya inostrannomu yazyku v vysshej shkole [Application of blended learning technologies in higher education] (In Russian). *Sovremennye nauchnye issledovaniya i innovatsii* [Modern Scientific Research and Innovation], 10. Retrieved December 29, 2021, from http://web.snauka.ru/issues/2013/10/27936
16. Maltsev, A. O. (2009). Sredstva kommunikatsij distantsionnogo obucheniya [Means of communication of distance learning] (In Russian). *Sovremennye problemy nauki i obrazovaniya* [Modern Problems of Science and Education], 3. Retrieved May 03, 2022, from https://science-education.ru/ru/article/view?id=2327
17. Manapova, O. N. (2021). Sovremennye messendzhery v uchebnom protsesse professional'noj obrazovatel'noj organizatsii: sil'nye i slabye storony [Modern messengers in the educational process of a professional educational organization: strengths and weaknesses] (In Russian). *Innovatsionnoe razvitie professional'nogo obrazovaniya* [Innovative Development of Professional Education], *3*(31), 54–59.
18. Nazarenko, A. L., Rawson-Jones, K., & Anoshkina, J. G. (2004). Distantsionnoe obrazovanie i prepodavanie inostrannykh yazykov (opyt fakul'teta inostrannykh yazykov MGU) [Distance education and foreign language teaching (Experience of the Faculty of Foreign Languages,

Moscow State University)] (In Russian). Vestnik Moskovskogo universiteta. Seriya 19: Lingvistika i mezhkul'turnaya kommunikatsiya [Bulletin of Moscow University. Series 19: Linguistics and Intercultural Communication], Izd-vo Moskovskogo universiteta, 1, 205–210.

19. Patrusheva, L. S. (2014). Forum kak internet-resurs na urokakh RKI [Forum as an Internet resource in lessons of Russian as a foreign language] (In Russian). In III Mezhdunarodnyj nauchno-metodicheskij distantsionnyj seminar «Prepodavanie russkogo yazyka kak inostrannogo: teoriya i praktika»: Tomskij gosudarstvennyj pedagogicheskij universitet (Rossiya), fond «Russko-pol'skij institut» (g. Vrotslav, Pol'sha) [III International Scientific and Methodical Distance Seminar "Teaching Russian as a Foreign Language: Theory and Practice": Tomsk State Pedagogical University (Russia), Foundation "Russian-Polish Institute" (Wroclaw, Poland)], TGPU. Retrieved March 25, 2022, from http://tspu.edu.ru/images/uchim.russkij/folder1/семинар/Patrysheva.pdf
20. Pyannikov, M. M. (2010). K voprosu o ponyatiyakh «distantsionnoe obuchenie» i «distantsionnoe obrazovanie» [On the concepts of "Distance Learning" and "Distance Education"] (In Russian). *Gumanitarnyj vector* [Humanitarian Vector], *1*, 41–45.
21. Robert, I. V. (2010). *Sovremennye informatsionnye tekhnologii v obrazovanii: Didakticheskie problemy; perspektivy ispol'zovaniya* [Modern information technologies in education: Didactic problems; prospects for use] (In Russian). IIO RAO.
22. Romanova, S. M. (2013). Sistema distantsionnogo obucheniya kak sredstvo informatsionno-kommunikatsionnykh tekhnologij v obrazovatel'nom protsesse [Distance learning system as a means of information and communication technologies in the educational process] (In Russian). *Nauchno-metodicheskij ehlektronnyj zhurnal «Kontsept»* [Scientific-Methodical Electronic Journal "Concept"], *4*, 271–275. Retrieved April 02, 2022, from http://e-koncept.ru/2013/64056
23. Savilova, S. L., Kropotkina, A. A., Kokhanovskaya, E. V., Smychkova, E. G., & Chaj, M. A. (2020). Distantsionnoe obuchenie inostrannomu yazyku v period pandemii na primere russkogo yazyka kak inostrannogo: iz opyta raboty [Distance learning in a foreign language during the pandemic on the example of Russian as a foreign language: from experience] (In Russian). In Tsifrovaya gumanitaristika i tekhnologii v obrazovanii (DHTE 2020): sb. materialov Vserossijskoj nauchno-prakticheskoj konferentsii s mezhdunarodnym uchastiem. 19–21 noyabrya 2020 g. [Digital Humanitarianism and Technologies in Education (DHTE 2020): Proceedings of the All-Russian Scientific-Practical Conference with International Participation. November 19–21, 2020] (pp. 112–123). Izdatel'stvo FGBOU VO MGPPU, Moscow.
24. Savostyanova, Y. I., & Sichinava, Y. N. (2014). Integratsiya sistemy yazykovykh podkastov v uchebnyj protsess voennogo vuza [Integration of the system of language podcasts in the educational process of a military university] (In Russian). *Teoriya i praktika obshhestvennogo razvitiya* [Theory and Practice of Social Development], *14*, 35–38. Retrieved November 09, 2021, from https://readera.org/14932380/
25. Senchenkova, E. V. (2016). Interaktivnye formy obucheniya russkomu yazyku kak inostrannomu v meditsinskom vuze [Interactive forms of teaching Russian as a foreign language in a medical school] (In Russian). *Pedagogicheskoe obrazovanie v Rossii* [Pedagogical Education in Russia], *11*, 96–99.
26. Shumilin, V. P., & Shumilina, N. G. (2019). Ispol'zovanie internet-tekhnologij v obrazovanii [Using Internet technologies in education] (In Russian). *Uchenye zapiski Orlovskogo gosudarstvennogo universiteta* [Scientific Notes of Orel State University], *1*(82), 355–357.
27. Sysoev, P. V. (2014). Podkasty v obuchenii inostrannomu yazyku [Podcasts in Teaching a Foreign Language] (In Russian). *Yazyk i kul'tura* [Language and Culture], *2* (26), 189–201.
28. Tseryulnik, A. Yu. (2020). Ispol'zovanie distantsionnogo formata obucheniya studentov v obrazovatel'nom protsesse [The use of distance learning format for students in the educational process] (In Russian). *Mezhdunarodnyj nauchno-issledovatel'skij zhurnal* [International Research Journal], *6*(96), part. 3, 92–95.

29. Uman, A. I., & Fedorova, M. A. (2017). Uchebnoe zadanie kak sredstvo formirovaniya uchebnoj samostoyatel'noj deyatel'nosti [Learning task as a means of independent learning activities formation] (In Russian). *Problemy sovremennogo obrazovaniya* [Problems of Modern Education], 2, 111–117. Retrieved April 29, 2022, from http://www.pmedu.ru
30. Volkova, A. N. (2001). Spetsifika yazykovogo distantsionnogo obucheniya [Specifics of language distance learning] (In Russian). In *Inostrannye yazyki v vysshej shkole: problemy, opyt, perspektivy: materialy mezhvuzovskogo nauchno-metodicheskogo seminara.* Chita. 13 marta 2001 g. [Foreign Languages in Higher Education: Problems, Experience, Prospects: Proceedings of the Interuniversity Scientific and Methodological Seminar. Chita. March 13, 2001] (pp. 30–35). ZIP Sib UPK, Chita.
31. Vyazovskaya, V. V., Danilevskaya, T. A., & Trubchaninova, M. E. (2020). Internet-resursy v obuchenii russkomu yazyku kak inostrannomu: ozhidaniya vs real'nost' [Internet resources in teaching Russian as a foreign language: expectations vs reality] (In Russian). *Rusistika* [Russian Language Studies], *18*(1), 69–84.

Machine Translation Versus Human Translation of Artionyms

Natalia Shutemova

Abstract The paper considers how artionyms are represented in human and machine translation, which is relevant for applied, corpus and comparative linguistics, as well as Translation Studies. Names of works of art form national and world artionymicon, participate in art discourse, influence the quality of its translation and the evaluation of chef-d'oeuvres in target cultures, which determines the research relevance. The notion "artyonym" is considered in reference to the notion of essence applied to typological properties of masterpieces, including their ideas, emotivity, imagery and artistic form. The comparative analysis of source and target artionyms examplified by English translation of K. Malevich's artionyms has revealed that human translation is characterized with "sense-for-sense" creative representation of source artionyms in the target language and culture and is based on comprehending an artionym in the context with the form and content of the work of art itself. Machine translation is devoid of contextualization and implements the tactics of "word-for-word" automatic reproduction. The classification of target artionyms coincidence in human and machine translation includes full, partial and noncoincidence influencing the evaluation of Russian fine arts of the twentieth century abroad.

Keywords Art discourse · Artionym · Machine translation · Smart technologies in discourse analysis · Tagging for corpus analysis

1 Introduction

Accuracy in translation is one of the most relevant issues in Translation Studies, and it has been considered from various perspectives due to its multidimensional character. On the one hand, it is regarded as accuracy of representing semantics of lexical units and their stylistic connotations. On the other hand, it deals with challenges of representing the text as a whole comprising its form and content. It is important that both aspects of the problem influence the quality of translation.

N. Shutemova (✉)
St. Petersburg University, Universitetskaya Emb. 7/9, 199034 St. Petersburg, Russia
e-mail: n.shutemova@spbu.ru

P. Eismont et al. (eds.), *Literature, Language and Computing*,
https://doi.org/10.1007/978-981-99-3604-5_17

Difficulties of translation are determined by heterogeneity of minds, languages, cultures, but may be overcome based on common reality, ability to analyze and understand the essence using logic procedures. However, these factors have led scholars to the idea that precision stipulated as a target is illusionary, utopic in translation [1–3]. This thesis is represented, e.g., in research of interference and conceptual mismatches in translation of terminology and allusions [4–6]. In the contemporary digital paradigm in linguistics [7, 8], accuracy in translation has become even more acute for Machine Translation (MT) and Computer Assisted Translation (CAT) [9].

These tools are being increasingly used nowadays, and their analysis has shown they may be efficiently applied for translating texts which type is called "rigid". However, for translating "soft" texts they are thought to be inefficient. The former group includes such texts as instructions and documents, e.g., contracts, letters of credit, bills of lading, while the latter—literary texts and advertisement. This difference is explained by standard lexicon, syntax and composition of the former, in contrast to the creative, experimental nature of the latter.

With the development of world culture and tourism, MT and CAT have been more intensely applied to translate art discourse, in particular booklets, exhibition catalogues, audio guides aimed at mass target group. At the same time, the necessity to use MT and CAT in translation projects in this field has evoked demads for strategies to analyze strengths and weaknesses of this practice. Specific attention is paid to translation errors, tactics of revealing, classifying and fixing them. For instance, one can distinguish general and specific errors in MT and CAT: general errors include shifts in cohesion, coherence and style, while specific errors deal with punctuation, spelling, anaphora and are caused by limits of the computer interface, a sentence bounded strategy, interference of a native language [10].

The quality of translation in the field of arts and Arts Studies depends on various factors, one of them being the accuracy of representing artionyms of the source text (ST) in the target culture (TC). This accuracy influences the extent to which the information conveyed by the ST is represented in the target text (TT). Hence, it influences the quality of translation and the evaluation of works of art in foreign cultures. The purpose of this paper is to consider how artionyms are translated by highly available and mostly applied MT tools, such as "Google Translator", "Yandex Translator" in comparison with human translation.

2 Knowledge Background

Preceding the analysis of the subject matter it is necessary to mention that artionyms are traditionally defined as names of works of art. Hence, they are regarded as proper nouns having special semantics not limited to words they comprise, and a special structure that can be modeled by means of formulae. The research of artionyms in semantic, structural and functional aspects has evolved a system of interconnected notions. Thus, a branch of onomastics studying names of works of art is called artionymics. To denote a scope of artionyms of a group of people/nation in a specific period the term “artionymy” is used. Moreover, the term “artionymicon” has been coined to refer to a structurally organized vocabulary consisting of artionyms and containing information about authors of works of art, their dates, aesthetic trends they belong to. In its turn, an artionym system differs from the artionymicon and artionymy as far as it is treated as a coherent and structured scope in which artionyms are organized in oppositions. A process of transforming a lexical unit into an artionym is nominated “artionymization”. An artionym model is also considered and defined as a scheme showing artionym structure and derivation. Besides, the term “artionym formula” is used to mean a sequence of components in the name of a work of art. The classification of artionyms includes topoartionyms and anthropoartionyms, the former being names of works of art derived from toponyms and the latter being names of works of art derived from anthroponyms. The text is characterized as artionymic if it frequently uses artionyms and allows to analyze principles of their derivation and functionality [11].

We believe that artionyms, as names of works of art, function as dominant words representing essence of their artworks.

- Firstly, it means that they express a scope of the author’s ideas generated during the process of artistic cognition of the world, the man and their interrelation.
- Secondly, they verbalize feelings and emotions experienced by the author within this cognitive process.
- Thirdly, artionyms correlate with artistic images in which the author realizes his thoughts and emotional attitude towards the object of cognition.
- Fourthly, they are interconnected with a form of a work of art and represent its specific features.

Thus, artionyms comprise such essential features of any work of art as ideas, emotivity, imagery and artistic form. Therefore, they could be regarded as key words verbalizing maximum content in the minimum of form.

This paradox determines corresponding difficulties of artionyms translation because it should be aimed at representing the essence of a work of art in the target culture (TC), rather than be limited with a search for an equivalent word in the target language (TL). The latter approach regards artionyms translation as a simple process rather than a cognitive challenge caused by a cognitive dissonance experienced by the translator. This cognitive dissonance, as it was mentioned above, is rooted in heterogeneity of the following factors:

- the author’s and translator’s minds;
- the source and target languages;
- the source and target arts traditions;
- the source and target cultures.

Moreover, as far as the translation of artionyms deals with representing the essence of works of art in TC, and as far as, in its turn, the essence is a complicated category comprising a system of properties that differentiates this very object from all the others, it is acute to consider the following cognitive tasks of translating artionyms:

- to comprehend the essence of the work of art the name of which should be translated;
- to represent the essence of the work of art by means of an artionym in the target language.

The algorithm of solving the first task includes the following subtasks:

- to comprehend the specifics of the artistic form and to reveal its dominant characteristics;
- to analyze the imagery of the author's thought;
- to grasp the author's ideas and emotivity represented in the work of art;
- to form a cognitive model of the essence of the work of art;
- to understand how the essence of the work of art is interconnected with its title and to grasp the "depth" of the latter.

The algorithm of solving the second task includes the following subtasks:

- to use the target language to represent the translator's mental model of the essence of the work of art including its ideas, emotivity, imagery and formal specifics;
- to represent the logic of interconnection between the essence of the work of art and its title;
- to construct the artionym in the target language.

Taking these cognitive aspects into account, we regard the process of translating artionyms in art discourse as a cognitive challenge, non-standard task requiring creative decision-making. It is obvious that artificial intelligence is unable to follow this cognitive algorithm, however, one can observe coincidence in machine and human variants of artionyms translation.

3 Results

Our consideration of machine and human translation of artionyms is based on comparative analysis of artionyms and their translation variants collected by ourselves into a parallel corpus for the purposes of this research. Following traditions in terminology of Translation Studies authentic artionyms are named as source artionyms (SA), while their translation variants as target artionyms (TA). In this paper source artionyms are considered in the context of Russian visual arts of the twentieth century, in particular works by K. Malevich, exhibited in the State Russian Museum, Saint Petersburg. Target artionyms are represented by English translation of Russian artionyms and include four groups:

- variants of human translation of names of works by K. Malevich that are officially adopted and used in the museum, they are available at the museum website [12];

- target artionyms generated by "Yandex Translator";
- target artionyms generated by "Google Translator".

The procedure of constructing the corpus included several steps:

- "field" study of data, which means that source artionyms were collected in the museum itself because it allowed us to analyze artionyms in their connection with authentic works of art, rather than to borrow them mechanically without comprehending the essence of K. Malevich's masterpieces;
- "field" study of target artionyms that are translated by a human (HTA) and adopted in the museum;
- collection of TAs generated by machine translation (MTA);
- comparison of HTAs and MTAs.

Thus, the parallel corpus of SAs and TAs includes K. Malevich's artionyms as source units in Russian and three variants of target units in English.

The comparative analysis of SAs and TAs has revealed the following trends:

- all TAs represent a strategy of reconstructing SAs;
- results of human and machine translation of artionyms do not always coincide with each other;
- results of human and machine translation of artionyms may coincide with each other;
- results of machine translation of artionyms may coincide with each other;
- results of machine translation of artionyms do not always coincide with each other;
- tactics of human translation of artionyms represent cognitive processes of interpreting the source idea, emotivity and imagery;
- tactics of machine translation of artionyms represent an automatic "word-for-word" approach and include substitution.

The strategy of reconstructing SA in human translation is cognitive in its nature: it means that the translator comprehends the connection between SA and ideas, emotivity and imagery of the work of art, forms a cognitive model of this connection in the translator's mind, represents this model in TL so that the SA semantics and form are represented in TA in full. However, the strategy of reconstructing TA in machine translation is not cognitive: it is based on substituting a word in SL with a word in TL rather than understanding the genesis of SA or its cognitive modeling. In general, results of human and machine translation of artionyms can be classified into three types of coincidence of HTA with MTA: complete coincidence, partial coincidence, noncoincidence, that will be considered herein.

Full coincidence

The first type of correlation between HTA and MTA may be characterized as full coincidence occurring in case SAs are brief, precise and they clearly nominate images depicted on canvases. For example, names of the world-known masterpieces "Cherny kvadrat", "Cherny krug", "Cherny krest" are represented by the following

HTAs: “Black Square”, “Black Circle”, “Black Cross”, which almost completely match corresponding MTAs (aside from initial capital letters in HTAs and lowercase letters in MTAs).

The noun “pejzazh” (“landscape”) is a frequently used term denoting a genre in fine arts and is applied to nominate an individual work of art belonging to this genre. A regular means of representing it in English is “landscape” which is used similarly to denote both a genre and a work of art. Hence, full coincidence of the HTA and MTA is quite predictable and of no wonder. It is this very term that is used both in HT and MT. The same can be mentioned concerning the noun “portrait” which is used to name a work of art depending on its belonging to a certain genre in fine arts. For example, one of its kinds is termed “self-portrait” both in HT and MT.

The comparative analysis of SAs and TAs reveals that SAs containing words borrowed from foreign languages can be translated by means of source units, which results in coinciding HTA and MTA. For example, the SA “Aviator” is represented as “Aviator” in HT and MT, unlike the SA “Stroitel” represented as “Builder”, or “Kuznets” translated as “Blacksmith”, or “Krest’yane” rendered as “Peasants”.

A wide range of examples illustrating coincidence of MTs with HTs is given in Table 1, where the first column contains SAs, the second one—HTAs, while the third column—MTAs.

The analysis of these data reveals the only difference between HTAs and MTAs provided in the table above. It deals with capital letters that should be used in English titles according to spelling norms of the English language. However, this difference is quite illustrative because it shows that in MT there is no understanding of the background, content and purpose of translation. In other words, MT tools do not “understand” that the task deals with artionyms, it neither “recognizes” them nor correlates words and their combinations with real works of art, unlike human translation.

Partial Coincidence and Noncoincidence

The second type of correlation between HTA and MTA, that can be characterized as partial, takes place when HTA and MTA contain some common components. In this case MT processes SAs automatically and provides their “word-for-word” translation, while HT demonstrates traces of their cognitive processing. Based on the comparative analysis of HTAs and MTAs, discrepancies between them can be revealed in the following aspects: lexical semantic, grammar, stylistic, structural.

Lexical semantic aspect. The SA “Krasnyj kvadrat (Zhivopisnyj realizm krest’yanki v dvuh izmereniyah)” is represented by the HTA “Red Square. Painterly Realism of a Peasant Woman in Two Dimensions” which differs from the MTA “Red Square (Picturesque realism of a peasant woman in two dimensions)”. The adjective “picturesque” used in MTA means “attractive in appearance” [13] and changes the idea expressed by the SA where the adjective “painterly” refers to the key quality of painting and the ability to work with colour rather than to attractive appearance.

The SA “Otdyh (Obshchestvo v tsilindrah)” is represented by HTA “Relaxing (Society in Top Hats)”, which differs from MTA “Recreation (Society in top hats)” generated by “Yandex Translator” and from MTA “Leisure (Society in top hats)”

Table 1 HTA versus MTA: coincidence

SA	HTA	MTA (Yandex & Google)
Pejzazh	Landscape	Landscape
Pejzazh s zheltym domom (Zimnij pejzazh)	Landscape with a yellow house (Winter Landscape)	Landscape with a yellow house (Winter Landscape)
Avtoportret	Self-Portrait	Self-portrait
Molitva	Prayer	Prayer
Stroitel’	Builder	Builder
Aviator	Aviator	Aviator
Suprematism	Suprematism	Suprematism
Suprematism (supremus № 56)	Suprematism (Supremus № 56)	Suprematism (supremus № 56)
Chajnik s kryshkoj	Kettle with lid	Kettle with lid
Pejzazh s pyat’yu domami	Landscape with five houses	Landscape with five houses
Dve muzhskie figury		Two male figures
Krest’yane	Peasants	Peasants
Krest’yanka	Peasant woman	Peasant woman
Krest’yanin	Peasant	Peasant
Tri zhenskie figury		Three female figures
Krasnaya konnitsa	Red cavalry	Red cavalry
Krasnyj dom	Red house	Red house
Avtoportret (hudozhnik)	Self-Portrait (Artist)	Self-portrait (Artist)
Kuznets	Blacksmith	Blacksmith
Muzhskoj portret	Male portrait	Male portrait
Portret zheny hudozhnika	Portrait of the Artist’s wife	Portrait of the artist’s wife
Trojnoj portret	Triple portrait	Triple portrait
Tors (Figura s rozovym litsom)	Torso (Figure with a Pink Face)	Torso (Figure with a pink face)

constructed by “Google Translator”. Although the second component of SA given in brackets is represented in HTA and MTA similarly, the first component is interpreted differently: if HTA emphasizes the continuality of the process, MTAs accentuate the state of rest. Similar cases are summarized in Table 2.

The HTA of the SA “Portret Ivana Vasil’evicha Klyuna (Usovershenstvovannyj portret Klyuna)” is noteworthy because it allows one to trace the cognitive process of compression and demonstrates “sense-for-sense” approach, rather than “word-for word” used in MT. In HTA two components of the SA are merged, while in MTAs they are kept. Moreover, in HT and MT the idea is represented by means of lexical units “perfected” and “improved” correspondingly, that are close but still different in meaning. Despite common components in their semantics, the adjective “perfected” derived from the verb “to perfect” meaning “to make something free from faults” [13]

Table 2 Lexical partial coincidence

SA	HTA	MTA
Otdyh (Obshchestvo v tsilindrah)	Relaxing (Society in Top Hats)	Recreation (Society in top hats)/Leisure (Society in top hats)
Krasnyj kvadrat (Zhivopisnyj realizm krest'yanki v dvuh izmereniyah)	Red square. Painterly realism of a peasant woman in two dimensions	Red square (Picturesque realism of a peasant woman in two dimensions)
Portret Ivana Vasil'evicha Klyuna (Usovershenstvovannyj portret Klyuna)	Perfected portrait of Ivan Kliun	Portrait of Ivan Vasilyevich Klyun (Improved portrait of Klyun)/Portrait of Ivan Vasilyevich Klyun (Improved portrait of Klyun)
Kompozitsiya s Dzhokondoj («Chastichnoe zatmenie»)	Composition with La Gioconda (Partial Eclipse)	Composition with Gioconda ("Partial Eclipse")/ Composition with Mona Lisa ("Partial Eclipse")
Zhatva	Harvesting	Harvest
Devushka v derevne	Girl in the countryside	A girl in the village/Girl in the village
Yablonya v tsvetu	Apple tree in blossom	Apple tree in bloom
Na zhatvu (Marfa i Van'ka)	At harvesting (Marfa and Vanka)	To the harvest (Marfa and Vanka)/Harvest (Martha and Vanka)
Tors (Pervoobrazovanie novogo obraza)	Torso (Prototype of a New Image)	Torso (Initial formation of a new image)/Torso (Protogenesis of a new image)
Portret yunoshi	Portrait of a youth	Portrait of a young man

represents the idea of the portrait that the artist made "complete and correct in every way, of the best possible type or without fault" [13], while the adjective "improved" derived from the verb "to improve" expresses the idea of "getting better" [13]. In comparison with the participle "usovershenstvovannyj" in the SA, the HTA is more accurate than the MTA.

It is quite interesting that HTA and MTA for the SA "Yablonya v tsvetu" do not coincide either. To represent the imagery of K Malevich's thought, the noun "blossom" is used in the HTA, while in the MTA the noun "tsvet" is substituted with the noun "bloom", with both being semanticaly interconnected. "Blossom" denotes "a small flower, or the small flowers on a tree or plant", while "bloom" means "a flower on a plant" [13]. Both nouns can be used in idioms: "to be in blossom", "to be in bloom", with both signifying the state of "having flowers growing" and "producing flowers" [13] correspondingly. Thus, the HTA "Apple Tree in Blossom" accentuates the image of an apple tree shrouded in the host of small flowers, while the MTA "Apple Tree in Bloom" does not highlight the "smallness" of flowers.

Not less noteworthy is the SA "Na zhatvu (Marfa i Van'ka)" which representations in HTA and MTAs do not match. All three variants of translation use the lexical unit "harvest" denoting "the time of year when crops are cut and collected from the fields, or the activity of cutting and collecting them, or the crops that are cut or collected" [CD], which correlates with the meaning of the noun "zhatva" in the SA. However, in the HTA "At Harvesting (Marfa and Vanka)" the gerund derived from the verb "to harvest" is used, which accentuates the dynamic aspect of the image. At the same time in reference to the depicted image the preposition "at" in combination with the gerund can succinctly actualize meanings referring to time, place, activity and direction expressed by the SA. Hence, the HTA can be interpreted in various aspects including both dynamics and direction. In contrast to the HTA, the MTA generated by "Yandex Translator" "To the harvest (Marfa and Vanka)" represents "word-for-word" approach clearly expressing the idea of direction by means of the preposition "to" without emphasizing dynamics. The variant "Harvest (Martha and Vanka)" provided by "Google Translator" and using the noun reduces the dynamic aspect of the imagery and the idea of movement, process, direction represented in the SA.

Grammar aspect. Based on the comparative analysis of SAs, HAs and TAs, one can differentiate two main kinds of mismatches: word order and articles, that are summarized in Table 3.

For example, shifts in word order can be demonstrated by means of the SA "Golova krest'yanina". It is represented by the HTA containing the so-called "of-phrase" (a noun with the preposition "of"). In contrast, both MTAs use the possessive case which is more informal. Moreover, their structure accentuates the noun "head", while word order in the HTA accentuates both nouns similarly to the SA.

The word order shift is observed in variants of translating the SA "Slozhnoe predchuvstvie (Tors v zheltoj rubashke)" as well. The HTA "Torso in a Yellow Shirt (Complicated Premonition)" contains the word combination with a right-hand attribute, which is also used in the MTA "A complicated premonition (Torso in a yellow shirt)" generated by "Yandex Translator". However, in the MTA "Complicated Premonition (Yellow Shirt Torso)" generated by "Google Translator" includes the word combination with a left-hand attribute.

Table 3 Grammar partial coincidence

SA	HTA	MTA
Golova krest'yanina	Head of a peasant	The peasant's head/Peasant's head
Slozhnoe predchuvstvie (Tors v zheltoj rubashke)	Torso in a yellow shirt (Complicated Premonition)	A complicated premonition (Torso in a yellow shirt)/Complicated Premonition (Yellow Shirt Torso)
Devushki v pole	Girls in a field	Girls in the field
Krest'yanin v pole	Peasant in a field	A peasant in the field/Peasant in the field
Plotnik	Carpenter	A carpenter

Concerning definite and indefinite articles HTAs demonstrate a consistent approach according to which initial main nouns in SAs are used without articles (“Head”, “Torso”, “Girls”, “Peasant”, “Carpenter”), which has a generalizing, abstract meaning and represents more general, abstract ideas rather than individual objects (definite or indefinite). Subsequent class-nouns in HTAs are consistenly used with indefinite articles without indicating definite objects distinct from all other objects of a certain class, hence, in a more general and abstract meaning rather than particularizing. In contrast, MTAs represent inconsistent approach in using articles. Moreover, the use of the definite article with class-nouns (e.g., “the field”) limits and narrows their meanings, distinguishing a certain object, making it distinct from other objects. This feature contradicts the very philosophy and aesthetics of suprematism created by K. Malevich.

Stylistic aspect. The SA “Torzhestvo neba” is represented in the HTA “The Triumph of Heaven” which contains lofty nouns “triumph” and “heaven” because the semantics of the noun “torzhestvo” in the SA has a lofty connotation. Similar variant is generated by “Google Translator”; however, it lacks the definite article and capital initial letters. In comparison, MTA reconstructed by “Yandex Translator” demonstrates a “word-for-word” strategy: it includes the lofty noun “triumph” with the neutral noun “sky”.

The SA “Rabotnitsa” is represented with the HTA “Woman Worker” containing nouns that are stylistically neutral and frequently used. In contrast, the MTA “Employee” generated by “Yandex Translator” is formal in its stylistic connotation, as well as the MTA “Female worker” containing less frequently and more specifically used formal noun “female”.

Structural aspect. The comparative analysis has revealed shifts in structures of TAs summarized in Table 4.

On the one hand, in human translation the structure of SAs can be reversed. For example, the HTA “Torso in a Yellow Shirt (Complicated Premonition)” changes the sequence of logical parts of the SA “Slozhnoe predchuvstvie (Tors v zheltoj

Table 4 Structural partial coincidence

SA	HTA	MTA
Portret Ivana Vasil’evicha Klyuna (Usovershenstvovannyj portret Klyuna)	Perfected portrait of Ivan Kliun	Portrait of Ivan Vasilyevich Klyun (Improved portrait of Klyun)
Suprematizm (Supremus №58. Zheltoe i chernoe)	Suprematism (Supremus No. 58)	Suprematism (Supremus No. 58. Yellow and black)
Arhitekton “Al’fa”	Alpha. Architecton	Architecton “Alpha”
Arhitekton “Gota”	Gota. Architecton	Architecton “Gotha”
Slozhnoe predchuvstvie (Tors v zheltoj rubashke)	Torso in a yellow shirt (Complicated Premonition)	A complicated premonition (Torso in a yellow shirt)/ Complicated Premonition (Yellow Shirt Torso)

rubashke)", while both MTAs reproduce it automatically ("A complicated premonition (Torso in a yellow shirt)"/"Complicated Premonition (Yellow Shirt Torso)"). Similarly, the structure of the SAs "Arhitekton 'Al'fa'" and "Arhitekton 'Gota'" is reversed in corresponding HTAs: "Alpha. Architecton", "Gota. Architecton", while in corresponding MTAs it is reproduced "word-for-word": "Architecton 'Alpha'", "Architecton 'Gotha'".

On the other hand, in human translation the structure of SAs can be reduced. For example, the HTA "Perfected Portrait of Ivan Kliun" reduces one of the components of the complex SA "Portret Ivana Vasil'evicha Klyuna (Usovershenstvovannyj portret Klyuna)" both of which are automatically reproduced in identical MTAs "Portrait of Ivan Vasilyevich Klyun (Improved portrait of Klyun)". The second component of the SA "Suprematizm (Supremus № 58. Zheltoe i chernoe)" is also reduced in the HTA "Suprematism (Supremus No. 58)" and kept in identical MTAs "Suprematism (Supremus No. 58. Yellow and black)".

Finally, the comparative analysis shows that **the third type of correlation between HTA and MTA**, that can be characterized as noncoincidence, is also possible. As for works by K. Malevich, this type of correlation has been revealed only when the SA "Sportsmeny" is represented by the HTA "Sportsmen" and MTA "Athletes". However, it is not too rare in art history discourse where, e.g., the HTA "Abundance" differs from the MTA "Cornucopia".

4 Discussion and Conclusion

We suppose that the comparative research of artionyms and their human and machine translation is perspective in lexicology, Translation Studies, discourse, cultural, corpus and computational linguistics.

Firstly, the results may be used to design national special corpora of artionyms and texts that form art discourse. Such projects are helpful in outlining national artionymicon. Secondly, they may be applied to collect parallel corpus of source and target texts containing artionyms in various languages and to tag it either manually or automatically. Such corpus is thought to contribute to forming a multilingual system of artionyms helpful in their comparative study which, in its turn, is important for comparative study of languages, cultures, and mentalities. In other words, it could result in collecting international artionymicon and promote its comparative study.

Thirdly, the research can be taken into consideration for further enriching and updating Translation Memory, as well as improving the neural network machine translation systems and training them to recognise more precisely semantic components of artionyms or similar corpora-based terms and following the models designed by human translators. This will elaborate machine translation of art discourse, which could be applied for translating catalogues, booklets, audio-guides, exhibition advertisement. Moreover, the research results can be fruitfully introduced into didactics of both human and machine translation to improve overall accuracy and quality of human and machine translation.

Acknowledgements The reported study was funded by RFBR, project number 20-012-00276.

References

1. Cristinoi, A. (2016). Translation between typologically different languages or the Utopia of equivalence: 1 vs 1.round, 1.long or 1 nasty.being. In *Meaning in translation: Illusion of precision* (pp. 99–109). Cambridge Scholars Publishing, UK.
2. Kerremans, K. (2016). Illusion of terminological precision and consistency: A closer look at EU terminology and translation practices. In *Meaning in translation: Illusion of precision* (pp. 161–177). Cambridge Scholars Publishing, UK.
3. Filanti, R. (2016). "The Murder and the Echo": How meaning reverberates in translation. In *Meaning in translation: Illusion of precision* (pp. 321–333). Cambridge Scholars Publishing, UK.
4. Kwiatek, E. (2016). When terms do not match: Translation strategies for dealing with conceptual mismatches in surveying terminology. In *Meaning in translation: Illusion of precision* (pp. 237–251). Cambridge Scholars Publishing, UK.
5. Platonova, M. (2016). Striving for precision: Biblical allusions in terminology. In *Meaning in translation: Illusion of precision* (pp. 179–196). Cambridge Scholars Publishing, UK.
6. Temmerman, R. (2016). Translation and the dynamics of understanding words and terms in contexts. In *Meaning in translation: Illusion of precision* (pp. 139–161). Cambridge Scholars Publishing, UK.
7. Atkins, S., Clear, J., & Ostler, N. (1992). Corpus design criteria. *Literary and Linguistic Computing, 7*(1), 1–16. Oxford University Press. https://doi.org/10.1093/llo/7.1.1. Retrieved April 29, 2021, from https://www.researchgate.net/publication/242692098_Corpus_Design_Criteria
8. Biber, D. (1994). Representativeness in corpus design. In Zampolli, A., Calzolari, N., & Palmer, M. (Eds.), *Current issues in computational linguistics: In Honour of Don Walker*. Linguistica Computazionale (vol. 9, pp. 307–407). Springer. https://doi.org/10.1007/978-0-585-35958-8_20. Retrieved April 28, 2021, from https://link.springer.com/chapter/10.1007/978-0-585-35958-8_20?error=cookies_not_supported&code=6a8f92e0-34a1-4d58-9226-4f5568efaa0f
9. Qin, Y., Shang, J., & Lu, X. (2019). The gap between NMT and professional translation from the perspective of discourse. In *ACM international conference proceeding series. 3rd international conference on innovation in artificial* (pp. 50–54). Association for Computing Machinery. https://doi.org/10.1145/3319921.3319936. Retrieved May 27, 2021, from https://www.researchgate.net /publication/333183912_The_Gap_between_NMT_and_Professional_Translation_from_the _Perspective_of_Discourse
10. Ovchinnikova, I. G. (2019). Working on computer-assisted translation platforms: New advantages and new mistakes. *Russian Journal of Linguistics, 23*(2), 544−561. Moscow. https://doi.org/10.22363/2312-9182-2019-23-2-544-561. Retrieved May 02, 2022, from https://journals.rudn.ru/linguistics/article/view/21222/16991
11. Suprun, V. I. (2011). Razmyshleniya nad onomasticheskoj terminologiej [On ononmastics terminology]. Izvestiya Volgogradskogo Gosudarstvennogo Pedagogicheskogo Universiteta N 8 (62), 133–138. Volgograd.
12. http://en.rusmuseum.ru/. Retrieved May 03, 2022.
13. Cambridge Dictionary. Retrieved May 03, 2022, from https://dictionary.cambridge.org/

Computing in…… Emotional Text Processing

The Emotion in Text Analyzer: How to Visualize Its Output

Anastasia Kolmogorova and Alexander Kalinin

Abstract The article summarizes the results of the project conducted in the field of emotional text analysis. The project aim is to build up an analyzer able, according to the model of "Lövheim Cube", to detect eight emotions in the Internet-texts in Russian. Having collected a labeled dataset and trained ML models, we faced the problem of the way to visualize the results of emotional text analysis. The main problem is due to the non-discrete approach in emotion in text assessment we are following in our research: we consider emotions as continuum and not as discrete categories. In this paper, we briefly describe the main logic of our work on the analyzer and focus on the different approaches to its output visual presentation. Finally, we demonstrate the design we have chosen to fit users' expectations better and give an opening to the perspective of using the analyzer with such an interface.

Keywords Emotional text analysis · Computer analyzer · Visualization

1 Introduction

The article details the results of a project conducted in the field of Emotional Text Analysis—a rather recent but quickly developing path of the paradigm of Affective computing. The paradigm aim is to elaborate the technological tools, which will allow to the computers to detect humans' emotions, to propose an adequate response to them, to generate emotional output for influencing human user's emotional state.

In our research, we pursue the aim of creating an analyzer able to detect not only one leading emotion in text, but also eight basic emotions according to the emotion model of H. Lövheim.

One more advantage of our project is that it uses text data in Russian collected from the well-known social network VKontakte.

A. Kolmogorova (✉) · A. Kalinin
National Research University Higher School of Economics, Saint-Petersburg, Russian Federation
e-mail: akolmogorova@hse.ru

P. Eismont et al. (eds.), *Literature, Language and Computing*,
https://doi.org/10.1007/978-981-99-3604-5_18

At this stage of the research the crucial question that arises is: how do we visualize the results of emotional text processing done by machine learning models? As the analyzer is thought to be in use by a non-specialist in computer science or physiology of emotions, the interface should be friendly, convenient and with high usability level.

To be clear and avoid any confusion in exposing our content, we will follow the mentioned below logic: in Sect. 2 we will describe the main steps of elaborating the analyzer; Sect. 3 will focus on the analysis of related works in the field of data visualization; in Sect. 4 we'll detail the specificity of our design decision and in the Conclusions section, we will draw some perspectives of using our classifier equipped by the described interface in practice of professionals in different domains.

2 The Outlines of the Project

The concept of the emotion analyzer is based on an eight–membered model of emotions proposed by the Swedish researcher Lövheim [1]—the so-called «Lövheim Cube». Lövheim's initial hypothesis was that, although emotional states themselves are generated in the limbic system and amygdala of the brain, a further signal of emotion is activated and spreads to other parts of the brain due to the action of three monoamines: serotonin, dopamine and noradrenalin. Such a system of monoamine mediators serves as a kind of "emotional conduit" for transmitting information about emotions to all other parts of the brain. Later, Lövheim established the correlation of each of the eight basic emotions he had detected with a specific combination of the levels of three monoamines and visualized the model in the form of a cube on a coordinate plane with the axes 5-HT (serotonin), NE (norepinephrine), DA (dopamine).

Seven emotions in this classification have a double nomination, where the first part is the designation of the weakest degree of expression of the emotional state, and the second is the strongest (the exception is the emotion of Surprise): Interest/Excitement; Enjoyment/Joy; Surprise; Distress/Anguish; Anger/Rage; Fear/Terror; Contempt/Disgust; Shame/Humiliation.

In this concept, we were attracted by its two characteristics: a reasonable number of emotional classes and a heuristically promising way of visualizing the model itself.

A preliminary analysis of existing datasets in Russian has shown that the annotated collection of texts we need does not yet exist. This required the development of our own markup design, the search for a convenient platform for emotional annotation, the selection of an initial collection of texts for the formation of an annotated training set.

2.1 Dataset Retrieved for Annotation

First, before forming a corpus which could be submitted to annotation, we made a preselection of texts according to the following algorithm:

(1) expert monitoring of social networks to search for the most emotionally diverse and sufficient texts in terms of volume—we have selected the public groups “Overheard”, “Caramel”, “Room No. 6” of the Russian social network VKontakte;
(2) expert selection of hashtags correlated with each of eight basic emotions (for example, #_Fuu – for the emotion of disgust, #Sad or #Loneliness – for the emotion of Distress);
(3) validation of selected hashtags as markers of a group of emotional texts in a psycholinguistic experiment with 20 respondents;
(4) automatic extraction of content under validated hashtags (in total – 15,000 texts).

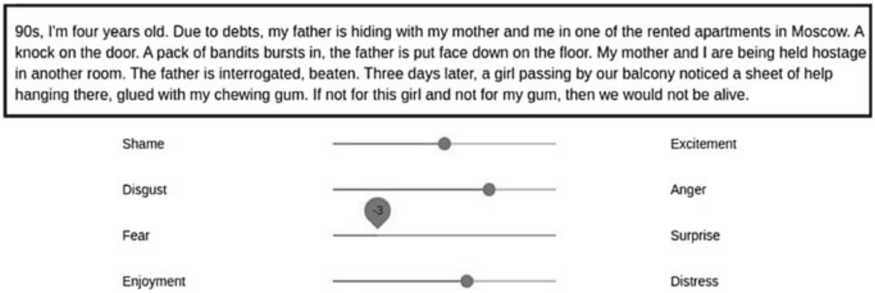
Afterwards, we made a random extraction of 400 texts from each of eight emotional text classes (3920 texts from 80 to 120 words). The mentioned 3920 texts built up the corpus for further emotional annotation.

2.2 Annotation Procedure

The assessment procedure itself was conducted on the Yandex-Toloka crowdsourcing platform, where informants with a rating of at least 90 percentile were selected, i.e. they proved themselves to be quite high-quality performers of crowdsourcing tasks. During the annotation process, the informants had to mark up the texts using the following scale and instructions.

Instructions: *Read the text carefully. If necessary, read it several times. What emotions does the author express in the text? On each of the scales, put a mark closer to the emotion that is more pronounced in the text. Put a mark as close as this emotion is obvious and strong in the text. For example, 1 step from the center—a shade of emotion is present, but weakly expressed; 3 steps—if the emotion is clearly present; 5 steps—if the emotion, without a doubt, dominates. If there are no emotions indicated on the scale in the text, leave a mark in the middle position.*

Next, the informants were offered four scales (Fig. 1), between the poles of which they could put a marker: Sadness—Enjoyment, Disgust—Anger, Shame—Excitement, Fear—Surprise. This design of scales is proposed because the “Lövheim Cube” model of emotions that we have taken as a basis, suggests that such emotions form oppositions at the biochemical level. For example, Sadness is “triggered” by low levels of serotonin and dopamine in the human blood, while Enjoyment is caused by high levels of these neurotransmitters. In the Cube model, these oppositions form its internal diagonals.



90s, I'm four years old. Due to debts, my father is hiding with my mother and me in one of the rented apartments in Moscow. A knock on the door. A pack of bandits bursts in, the father is put face down on the floor. My mother and I are being held hostage in another room. The father is interrogated, beaten. Three days later, a girl passing by our balcony noticed a sheet of help hanging there, glued with my chewing gum. If not for this girl and not for my gum, then we would not be alive.

Fig. 1 The annotation interface

The zero mark in the middle of the scale is the point “zero emotions”, the labels to the left of 0 are assigned, respectively, values from −1 to −5; the labels to the right of 0 are assigned values from 1 to 5. Afterwards, these vectors were averaged to obtain one resulting vector.

2.3 Applied Machine Learning Models

As our way of annotating texts dataset was non-discrete, we decided to persist in the same, non-discrete, logic and to run a number of regression models.

Working with regression models, we used identical training and test samples. The number of fragments for each of the eight classes of emotions in the training sample was as follows: Shame—757, Distress—551, Excitement—397, Joy—390, Anger—336, Disgust—307, Surprise—248, Fear—159.

The text data in the test sample is split in a similar way (the ratio of training and test is 80%/20%): texts marked as Shame—189, Distress—138, Inspiration—100, Joy—98, Anger—83, Disgust—76, Surprise—62, Fear—40.

A number of models were tested, but the best results were obtained by using the so-called transformer model Fine-tuned BERT that showed rather sufficient values of Mean Squared Error (3.59) and Mean Absolute Error (1.36).

3 Visualization in Data Analytics in Emotional Text Analysis

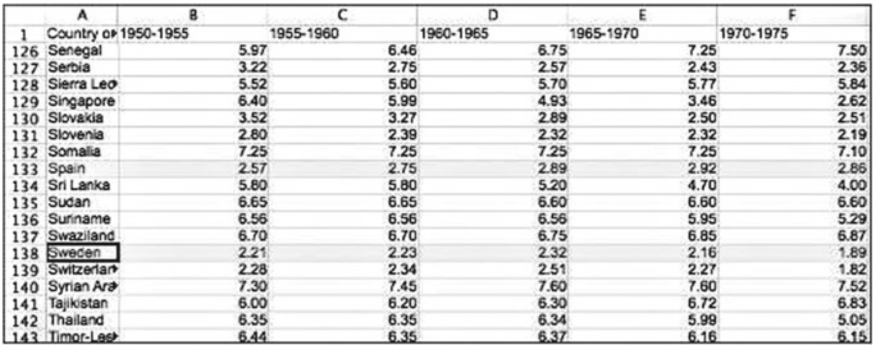
Data visualization plays an important role in the process of working with data, both from the side of studying patterns in samples, and for presentation purposes.

3.1 *The Role of Visualization in Data Analytics*

The reasons for the effectiveness of visual analysis tools are described in [2]. In particular, experiments indicate that visual information is processed in an optimal way and contributes to a faster decision-making process. The plausible explanation lies in the fact that the areas of the central nervous system responsible for processing visual signals have a longer phylogenetic history and therefore work faster [3]. Vision and the associated centers of the nervous system have played a crucial role in evolutionary development. Due to that even in modern humans' brain they provide a quick and accurate response to visual stimuli. The development of areas of the cerebral cortex and other departments responsible for processing semiotically complex structures (reading, calculating, understanding tables) has a later evolutionary age [Ibid]. This specificity in phylogenetic history explains the heuristic effectiveness of visual representation of data.

There are a significant number of approaches and principles that determine the choice of strategies for presenting information in visual form. One of the most authoritative approaches is the so–called "Visual Grammar", the ideas of which were developed initially in the classical work "Semiology of Graphics" [4], and then—in the Grammar of Graphics [5]. The main idea of the concept is to understand visualization as a function of the source data, usually presented in analytical (numerical form). In [6], the author demonstrates a perfect example of how differently we perceive and treat the same data given in a tabular format (Fig. 2) and in the form of a linear graph (Fig. 3).

The results of the conducted experimental study [Ibid] suggest that people who worked with visual data representation (Fig. 3) were faster to evaluate the dynamics in birth rate changes than those who worked with their tabular representation. Such a conclusion indicates the advantages of visual representation of data.



	A	B	C	D	E	F
1	Country o▸	1950-1955	1955-1960	1960-1965	1965-1970	1970-1975
126	Senegal	5.97	6.46	6.75	7.25	7.50
127	Serbia	3.22	2.75	2.57	2.43	2.36
128	Sierra Leo▸	5.52	5.60	5.70	5.77	5.84
129	Singapore	6.40	5.99	4.93	3.46	2.62
130	Slovakia	3.52	3.27	2.89	2.50	2.51
131	Slovenia	2.80	2.39	2.32	2.32	2.19
132	Somalia	7.25	7.25	7.25	7.25	7.10
133	Spain	2.57	2.75	2.89	2.92	2.86
134	Sri Lanka	5.80	5.80	5.20	4.70	4.00
135	Sudan	6.65	6.65	6.60	6.60	6.60
136	Suriname	6.56	6.56	6.56	5.95	5.29
137	Swaziland	6.70	6.70	6.75	6.85	6.87
138	Sweden	2.21	2.23	2.32	2.16	1.89
139	Switzerlar▸	2.28	2.34	2.51	2.27	1.82
140	Syrian Ara▸	7.30	7.45	7.60	7.60	7.52
141	Tajikistan	6.00	6.20	6.30	6.72	6.83
142	Thailand	6.35	6.35	6.34	5.99	5.05
143	Timor-Les▸	6.44	6.35	6.37	6.16	6.15

Fig. 2 Tabular representation of birth rate [Ibid]

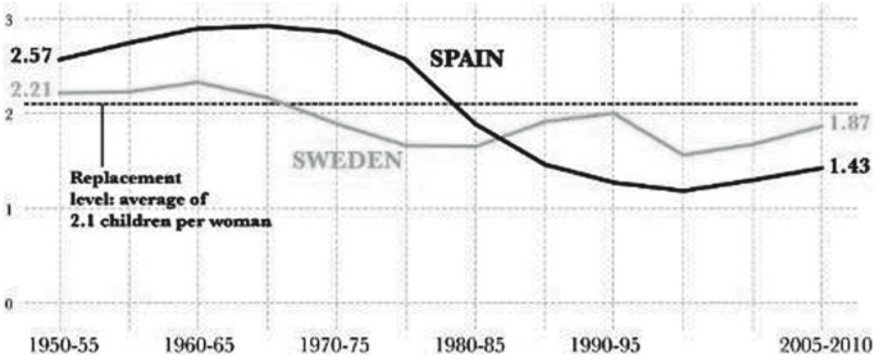


Fig. 3 Linear graph of birth rate [Ibid]

3.2 Visual Cues to Understand the Output of Sentiment/ Emotion in Text Analyzer

Like any other areas related to data representation, the tasks of tonality analysis and emotional analysis also require visualization tools. Their main function is to provide a "visual impression", which through clear visual keys will be able to recreate a generalized image of the process being studied or presented. The task of visualization in the analysis of emotions is to give a quick "emotional digest" of the text, to get an aggregated idea of the nature of the speaker's affective attitude to the events and phenomena mentioned in the text.

The simplest approach is to use bar charts. In [7], the visualization of sentiments from texts taken from Twitter and dedicated to the terrorist attacks on the Boston Marathon is presented in the form of pairs of bar charts. Each pair of columns represents the number of tokens related to a positive or negative sentiment. Within the framework of this approach, it is possible to estimate the proportion of emotionally colored words. The disadvantage of this kind of visualization is that it is impossible to determine the relative proportion of such words in the text and thus give a holistic visual image of an individual tweet.

The mentioned above problem was solved in another study by adding a neutral class and using stacked bar charts [8]. In this case, the sentiment is encoded in color (positive, neutral, negative), and the line width denotes the percentage of words of a certain sentiment. This approach is undoubtedly relevant, but its binary character imposes many limitations on the data presentation.

In [9], researchers studied about 132 papers on data visualization methods for solving problems of sentimental analysis of texts. According to the authors, the most common visual variable in studies on sentiment is color, and the most common task performed by an analyst is to assess the polarization of opinion. Visualization is most often used for a general overview or comparison. However, the analysis of emotions and affects, on the contrary, is little developed in the field of visualization. In the next Section, we will propose some techniques to fill the gap.

4 Different Models of Visualization of Emotion in Text Analysis: Advantages and Disadvantages

As the basic model in our project is a cube, it is easy to implement a spatial metaphor by itself as a visualization tool without any additional transformations. Within the model, emotional value of the text is considered as a point in the 3-dimensional space of this cube. An example of such visualization is shown in Fig. 4.

Such an approach, however, has two main drawbacks:

(1) 3-dimensional mapping. Data visualization, as a rule, uses the metaphor of a plane, not a three-dimensional space. This is due to the fact that a static display tool should be enough for us to quickly understand the static state. In the case of a 3-dimensional space, we will have to interact with the tool—rotate, change the angle, scale, etc., to clarify the position of the point. Undoubtedly, the use of interactive tools is justified for displaying dynamic processes, but for static characteristics the use of unnecessary clarifying actions is unacceptable, because it doesn't permit to achieve the main goal of such visualization—to give a quick idea of the phenomenon;

(2) the meaning of spatial axes is not quite obvious to the visualization client. Visualization should provide a person with a physical metaphor to understand the displayed processes. For example, an increase in the area is equivalent to the growth of some variable, a higher position indicates a higher order of the process under consideration. However, in the case of highly specialized concepts such as "dopamine level" or "serotonin level", such metaphors do not work, because it is not obvious to the visualization client (marketer, analyst, other specialist

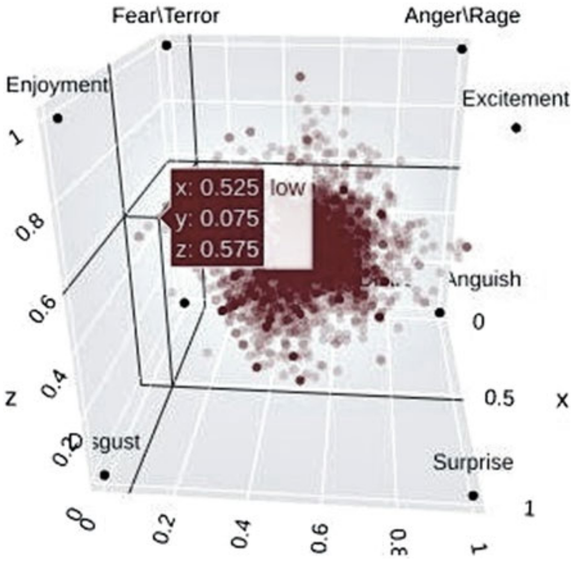


Fig. 4 The emotion in text evaluation given by the analyzer and visualized in form of a point located in the space of Lövheim cube

not involved in neurobiology) what changes and consequences an increase or decrease in the level of the neurotransmitter leads to. These processes are unfamiliar to non-specialists, unlike their innate abilities to determine physical processes and properties.

The next visualization concepts we tried were a bar chart (Fig. 5) and petal diagram (Fig. 6) models. In comparison with the previously tested approach, these two have many advantages, such as:

(1) diagonals used to assess emotion in text are scales (four scales) that can be depicted on a plane;
(2) due to the fact that these scales can be plotted on a plane, we do not need additional actions to clarify the configuration;
(3) scales are ordered sets of points, therefore, by the position of the points relative to each other, the degree of expression of a particular emotion can be compared;
(4) four axes enable multidimensional data representation.

Let us consider the application of this approach using the example of the following text, the evaluation of which we received as a result of markup on Yandex.Toloka:

(1) "*Besit, kogda taksisty otkruchivayut ruchki dlya podnyatiya stekol...poveliteli zhizni mlyayayaya*"
(“It pisses me off when taxi drivers unscrew the handles to raise the windows...the lords of life mlyayaya”)

As a result of the evaluation, this text received the following ratings on the scales: disgust_rage—4, fear_surprise—1, shame_excitation—−1, enjoyment_distress—1.

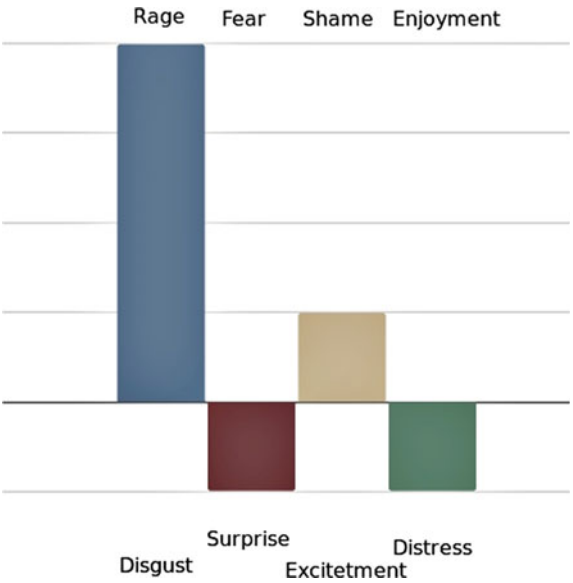


Fig. 5 Visualization of emotion in text assessment in the form of a bar chart

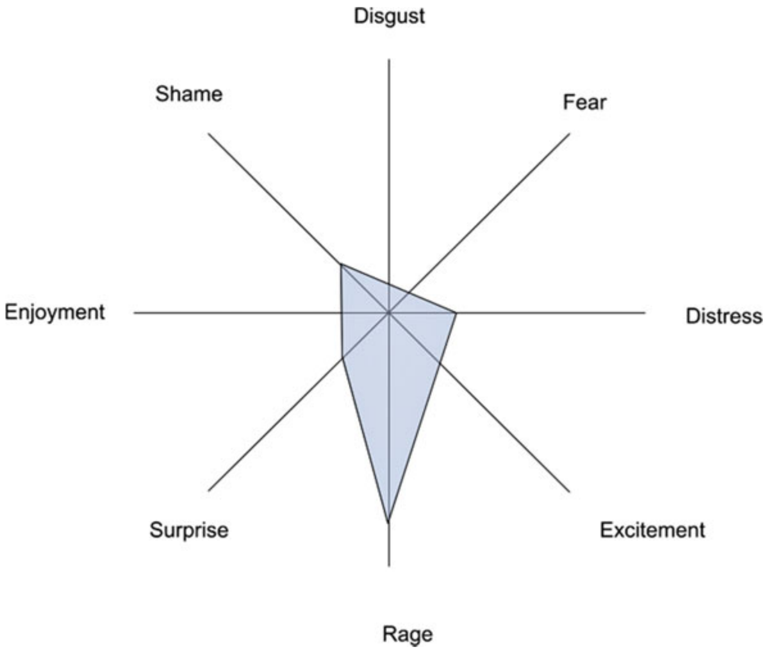


Fig. 6 Visualization of emotion in text assessment in the form of a petal diagram

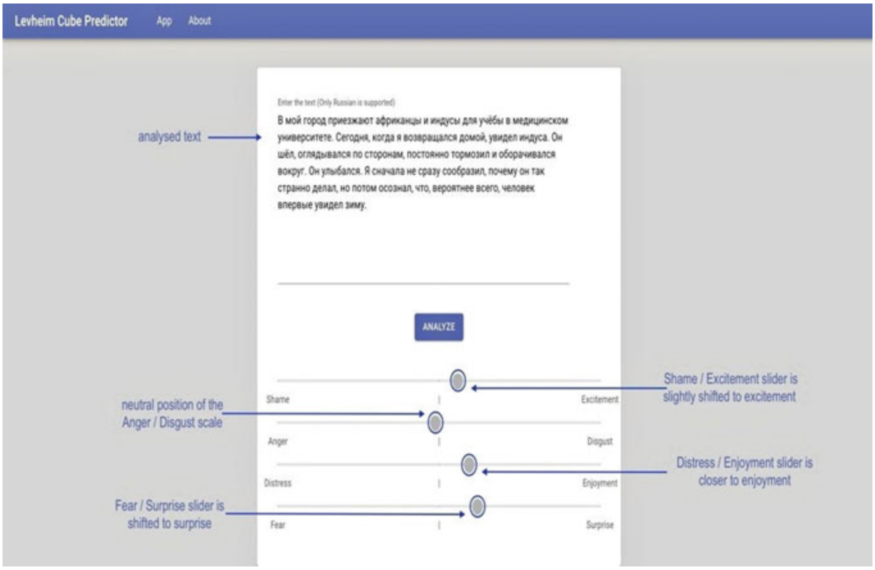
Both columnar and petal diagrams reflect more precisely the emotional configuration of the text, because they are isomorphic to the respondent's emotion assessment done while reading and interpreting the text.

For our third visual model we implemented an imitation of our interface for text emotional annotation: the user enters the text in a special field, clicks to the button and sees four sliders put differently on four scales according to the values of the emotional evaluation suggested by the analyzer (Fig. 7).

In particular, we see that the proposed text (Fig. 7) was evaluated by the program as a text in which the emotion of Shame is slightly expressed—the slider on the Shame/Excitement scale is a little shifted towards the Shame pole. At the same time, on the Anger/Disgust scale, the slider is in a neutral position—in the middle of the scale, at a point equivalent to zero emotions. On the Distress/Enjoyment scale, the slider is slightly put towards the Enjoyment pole. Finally, on the last, fourth scale (Fear/Surprise – Fear/Surprise) the slider is shifted towards Surprise.

5 Conclusions

In this paper, we demonstrated the necessity of using appropriate models for visualizing the results obtained in emotional texts processing.


Levheim Cube Predictor
App
About
analysed text
ANALYZE
Shame
Excitement
Anger
Disgust
Distress
Enjoyment
Fear
Surprise
Shame / Excitement slider is slightly shifted to excitement
neutral position of the Anger / Disgust scale
Distress / Enjoyment slider is closer to enjoyment
Fear / Surprise slider is shifted to surprise

Enter the text (Only Russian is supported)

В мой город приезжают африканцы и индусы для учёбы в медицинском университете. Сегодня, когда я возвращался домой, увидел индуса. Он шёл, оглядывался по сторонам, постоянно тормозил и оборачивался вокруг. Он улыбался. Я сначала не сразу сообразил, почему он так странно делал, но потом осознал, что, вероятнее всего, человек впервые увидел зиму.

Fig. 7 The interface of the emotion in the text analyzer in the form imitating the interface proposed for assessment

In the frame of our project conducted in the field of Emotional Text Analysis, we elaborated and presented four visual models plausibly compatible with our emotion in the text analyzer. After a critical analysis of them all, we should conclude that, now, the last among the discussed above models seems to fit best the conception of the non-discrete emotion assessment implemented in our research. However, we'll continue to search for the optimal way to provide a user with a quick emotional digest of a text.

At the same time, while designing the interface of the text analyzer, we should always be aware of the usability of the tool depending, in its turn, on the domains where it will find its application. Potentially, we see as such the following spheres:

- monitoring of social networks to detect publics or groups showing the patterns of destructive or aggressive behavior,
- permanent psychological assistance to the people with depression [10] or people with disabilities,
- conceiving text content for different didactic materials used in teaching Russian as a foreign language,
- monitoring of the subjective attitude towards goods and services offered by different companies,
- providing text users with more nuanced semantic information about a word being a part of word network, for example in Visual Thesaurus application.

Considered in these perspectives, the interface we propose, as a tool to analyze emotions in texts, should be intuitively clear, comprehensive and able to give a quick idea of emotion distribution in a text. The further work with different models will help us to assess the applicability of each of them for the task of visualizing results of emotion in the text analyzer.

References

1. Lövheim, H. (2012). A new three-dimensional model for emotions and monoamine neurotransmitters. *Medical Hypotheses, 78*, 341–348.
2. Priti, S. (1997). *A model of the cognitive and perceptual processes in graphical display comprehension* (AAAI Technical Report FS-97-03). Retrieved April 28, 2021, from https://www.aaai.org/Papers/Symposia/Fall/1997/FS-97-03/FS97-03-012.pdf
3. Filimon, F., Nelson, J. D., Hagler, D. J., & Sereno, M. I. (2007). Human cortical representations for reaching: Mirror neurons for execution, observation, and imagery. *NeuroImage, 37*(4), 1315–1328. https://doi.org/10.1016/j.neuroimage.2007.06.008
4. Bertine, J. (1967). *Semiology of graphics*. EsriPress.
5. Wilkinson, L. (1999). *The grammar of graphics*. Springer-Verlag.
6. Cairo, A. (2013). *Functional art*. New Riders.
7. Salarpour, A., Bamneshin, M., & Proios, D. (2013). *Sentiment analysis and visualization of social media data the #BostonMarathon #Bombings test case*. Retrieved April 28, 2021, from https://www.researchgate.net/publication/268503467_Sentiment_Analysis_and_Visualization_of_Social_Media_Data_The_BostonMarathon_Bombings_test_case
8. Visualization and sentiment analysis. Kaggle Notebook. Retrieved April 28, 2021, from https://www.kaggle.com/shaliniyaramada/visualization-and-sentiment-analysis
9. Kucher, K., Paradis, C., & Kerren, A. (2018). The state of the art in sentiment visualization. *Computer Graphics Forum, 37*, 71–96.
10. Huang, Y.-P., Goh, T. & Liew, C. L. (2007). Hunting suicide notes in web 2.0—Preliminary findings. In *ISMW '07: Ninth IEEE International Symposium On Multimedia—Workshops, Proceedings, IEEE Computer Society* (pp. 517–521). NW Washington.

The Multimedia Corpus of Russian Ironic Speech for Phonetic Analysis

Uliana Kochetkova, Pavel Skrelin, Vera Evdokimova, and Tatiana Kachkovskaia

Abstract This paper presents a detailed description of the multimedia corpus that was built for the phonetic analysis of Russian ironic speech. The corpus was developed after a series of preliminary studies. We analysed the expression of irony in films, series, audiobooks and other material presented in the open sources. Special attention was given to the contexts, in which the antiphrasis (irony as negation) appeared. Analysis of their lexical content, semantic structure and phonetic characteristics allowed us to construct the material for reading, which was the closest to the natural circumstances of irony expression. 60 Russian native speakers read the sets of short dialogues and coherent texts. Audio recording was simultaneously accomplished with video recording. The same target fragments were inserted in various types of sentences in ironic and non-ironic contexts. These homonymous fragments were extracted from the recorded material. Orthographic, prosodic and phonetic annotation was done. It included information about the context, target fragment transcription, stressed syllable and stressed vowel boundaries, intonation model, co-occurring paralinguistic phenomena and perceptual evaluation of irony by native listeners. A case study of phonetic properties of ironic speech showed that ironic utterances are characterized by the contrast in intensity level and stressed vowel duration as compared with non-ironic utterances, the different usage of intonation models in the two types of utterances was also observed.

Keywords Ironic speech · Phonetic characteristics · Paralinguistic phenomena

U. Kochetkova (✉) · P. Skrelin · V. Evdokimova · T. Kachkovskaia
St. Petersburg University, St. Petersburg, Russia
e-mail: u.kochetkova@spbu.ru

P. Skrelin
e-mail: p.skrelin@spbu.ru

V. Evdokimova
e-mail: v.evdokimova@spbu.ru

T. Kachkovskaia
e-mail: t.kachkovskaya@spbu.ru

P. Eismont et al. (eds.), *Literature, Language and Computing*,
https://doi.org/10.1007/978-981-99-3604-5_19

1 Introduction

The goal of the present paper is to describe the principles of construction and annotation of the multimedia corpus of Russian ironic speech. Detection of linguistic and paralinguistic cues of irony, including its acoustic characteristics, is one of the most challenging and urgent tasks nowadays, due to the high importance of the correct recognition of emotions, attitudes and connotations. Antiphrasis—the type of irony understood as negation of the lexical content by intonation or other means—is the phenomenon, which is crucial for understanding the real intention of the speaker and the correct meaning of the message during the dialogue. Without its true interpretation, the whole communication may fail.

In the second half of the twentieth century, various studies on irony started, considering it not only as a stylistic figure, but also and mostly as a cognitive phenomenon reflected in written and oral speech [1–3]. During the last fifteen years, numerous works appeared, in which the textual properties of irony are considered using automatic processing of big datasets [4–6]. Most of them investigate various kinds of written messages [6, 7–11]. During the last two decades, the amount of research using eye tracking techniques rose as well [12–14]. Multimedia corpora of ironic speech, including texts and audio recordings, were constructed mostly on the material of the English language and included audio recordings of the telephone dialogues, fragments of films and series from MTV, YouTube channels and other open sources [15].

In spite of the high interest in the analysis of irony in speech, its acoustic characteristics are still insufficiently described. Some fragmentary data may be found in the research of the last fifteen years [16–21], although some works date to the end of the twentieth century [22]. In the work of Kodzasov [23] the role of spectral characteristics and voice quality in irony expression in the Russian language is indicated, while other studies mostly consider the lexical and grammatical aspects of Russian ironic speech. Thus, the main purpose of our corpus is to allow detailed research on phonetic properties of irony in the Russian language.

At the same time, the role of gestures and mimics in speech cannot be underestimated [24]. A complex interaction between prosodic structure and paralinguistic phenomena was shown in a series of studies [25–28]. Therefore, in the present corpus, we included not only audio recordings, but also simultaneous video recordings. Such a multimodal corpus of ironic speech has not yet been built on the material of the Russian language.

2 Preliminary Research

Before the construction of the main part of the corpus, a preliminary research has been done. We selected ironic utterances from the audiobooks, films, series and TV shows available in open sources. The analysis of the ironic utterances in the actors' speech and of the contexts, in which they appeared, was highly important because these implementations of irony had already been approved by the directors (authors) of this audio or video production. A series of perceptual experiments provided data on the semantic structure, grammatical constructions, or lexical content of the utterances, in which the irony was well recognized. It helped elaborate material for the reading tasks. A description of the preliminary research was presented in the previous works [29–31].

3 Corpus Description

3.1 Reading Material

We constructed two types of texts: short texts or dialogues and long coherent texts. Each of the sets of short texts or dialogues (of 2–4 phrases) included from 60 to 82 short texts. Four long coherent texts were also created, each of them counting about 8000 characters. Both short and long texts included the homonymous target fragments, inserted in various types of sentences and in various contexts (with and without irony).

Short Texts and Dialogues. This type of material was composed in a way to allow the speaker to read as naturally as possible the target phrase. An example of the same target fragment in various contexts is given below:

- Ironic exclamatory sentence:
  - Chut’ ne utonula. Rusalochka! Ty zachem tuda poplyla? (Almost drowned. The little mermaid! Why did you go there?)
- Non-ironic narrative sentence:
  - —Kakoj tvoj lyubimyj personazh detskih skazok?—Rusalochka. (What is your favourite character in children’s fairy tales?—The little Mermaid.)
- Non-ironic exclamatory sentence:
  - Rusalochka! Kakoj krasivyj kostyum! Ty budesh’ luchshe vsekh! (The little Mermaid! What a beautiful costume! You’ll be the best!)

Coherent Texts. Each of the coherent texts included on the average 25 ironic fragments and 240 neutral, non-ironic utterances, because of the coherent narration requirements. There is an example extracted from one of such texts (ironic fragments are given in bold italics):

“…Koroleva skhvatila izyashchnymi dlinnymi pal’cami kolokol’chik i tri raza pozvonila v nego. Dveri raspahnul perepugannyj sluga, kotoryj sluchajno spotknulsya o dubovyj porog i upal, nadelav mnogo shumu.—Izvinite, ya…ya….—nachal on drozhashchim golosom.—***Izvinite***,—peredraznila ego Koroleva. Odnako sdelala ona eto bez privychnogo udovol'stviya,—***kakoj ty lovkij***! Vstavaj, da veli podavat' zavtrak v paradnuyu zalu! Zhivo!...".

(...The Queen grabbed the bell with her elegant long fingers and rang it three times. The doors were opened by a terrified servant who accidentally tripped over the oak threshold and fell, making a lot of noise. "I'm sorry, I... I..." he began in a trembling voice. "***I'm sorry***—" the Queen mimicked him. However, she did it without the usual pleasure—***how clever you are!*** Get up, and tell them to serve breakfast in the front hall! Now!...).

3.2 Experimental Design

During one recording session, speakers read one set of short texts and dialogues and one long coherent text. The total duration of one session of recording was no more than 40 min. The task was to read the mini-texts naturally, as they would implement such an utterance in an everyday talk.

During the second part of the task—the coherent text reading—the participants were supposed to read in an expressive way, but using their natural voice. Speakers were free to offer additional variants of ironic and non-ironic implementations when reading the short texts or long coherent texts.

3.3 Recording Conditions

The audio recordings were accomplished in a soundproof room using the audio interface Motu Traveller with the condenser cardioid microphone AKG HSC 271. The software included recording program Nuendo. All audio files were recorded in WAV format with 16-bit, 44,100 Hz sampling frequency. If the speaker consented to the video recording, it was done simultaneously with the audio recording on a Sony FDR-AX700 camcorder in a high-speed mode. Synchronization of the video recordings with audio will allow considering how linguistic and paralinguistic means accompany each other and what is the role they play in the resulting perception of irony.

3.4 Participants

The material was read by Russian native speakers (having standard pronunciation). The speakers did not have any special acting education or background. 60 speakers were recorded. Each of them accomplished at least one session of recording (i.e. 2 reading tasks: one set of short texts and dialogues and one of the long coherent texts). 27 men and 33 women took part in the experiment. Most part of the participants were

Table 1 Distribution of speakers in age groups

Age	Male	Female
17–25	17	21
25–35	4	5
Over 35	6	7

students from 17 to 25 years old; 9 people represented the middle age group—from 25 to 35; 13 participants were over the age of 35 (Table 1). The main part of this corpus of laboratory speech consists of the recordings of 10 speakers (5 men and 5 women) who read the entire material. In total, 15,690 stimuli were received (4520 ironic and 11,170 neutral stimuli).

4 Principles of Annotation

The main purpose, for which the corpus was built, is the potential study of acoustic cues of irony and its interplay with other linguistic and paralinguistic phenomena. Thus, the detailed orthographic, prosodic and phonetic annotation was needed.

4.1 Annotation Within the File

Two types of file sets were created. The first set included whole contexts (one short text or dialogue in each of them) extracted from the read material. Within them the boundaries of the target fragment were indicated, orthographic transcription was done, stressed syllable in the target fragment was marked. Whole orthographic transcription with punctuation was given. This set of files will allow a deeper analysis of intonation in the neighbouring contexts and the interaction of their phonetic parameters with the characteristics of the target fragments.

The second set (the main part of the corpus) consists of target fragments only: ironic and non-ironic. It enables all types of fast data analysis using scripts or making manual search. Annotation of this set of files comprises the following tiers: context, target fragment, stressed syllable, stressed vowel, intonation pattern, paralinguistic phenomena and native listeners' evaluation. The fundamental tone frequency was calculated automatically in Praat and corrected manually using Wave Assistant software when needed.

Context. This tier represents the entire orthographic and syntactic transcription of the context, notably a short text or dialogue, or a part of a long coherent text, from which the target fragment was extracted. This tier is present in both sets of files.

Target Fragment. In the target fragment, orthographic annotation was done. Four experts defined the boundaries of the target fragments. These fragments always coincided with the intonation phrase. Thus, they were implemented within the interpausal units or coincided with the interpausal units. As the main goal of our corpus was to provide material for the comparison of the homonymous ironic and non-ironic target fragments and for the study of phonetic characteristics of these fragments, there was no need to indicate the interpausal unit boundaries in the material.

In some cases, lexical markers of irony were pronounced along with the target fragment within the same intonation phrase. But most of the target fragments didn't contain any marker or context. It allows conducting very accurate perceptual experiments, in which the participants rely only on phonetic (in case of auditory perception) or paralinguistic (in case of visual perception) characteristics. A series of such experiments, that we conducted using the part of our corpus, showed that both channels of perception can work separately and each of them can provide information about irony without any other marker (lexical or semantic) of irony.

Stressed Syllable. We manually annotated the stressed syllable (nucleus) of the intonation centre (focus) of the target fragment. If there were other words in the target fragment that were prosodically emphasised, we annotated stressed syllable in each of them. Two experts made decisions about the place of the intonation centre, as well as about the presence (absence) of the additional prosodically emphasised words.

Stressed Vowel. The stressed vowel was annotated manually within the stressed syllable of the intonation centre and words with additional prosodic emphasis, if any.

Intonation Pattern. This tier contains information about the intonation model implemented in the target fragment. For this purpose, we used the system of Russian intonation description suggested by Volskaya [32]. This system is an extension of the Bryzgunova's system [33] widely used when teaching Russian as a foreign language. It includes 13 major melodic types, each with up to 4 subtypes. The 13 major types are classified on the basis on both form (melodic pattern) and function (finality, prominence, question, exclamation, non-finality) [34]. As in many other languages, in Russian the same melodic pattern may accomplish different functions: e.g., the high-level tone (in terms of C. Ode, H*H H%) may be observed in exclamations as well as in emotional questions expressing disbelief or perplexity, and in non-final intonation phrases, including list items. All these cases would be labelled differently in terms of Volskaya's system, as it is assumed that a different function may result in different realization rules—intervals, timing, or interaction of melody with other prosodic features [34].

Paralinguistic Phenomena. There exist various approaches to the multimodal annotation [24, 35, 36]. Nevertheless, we did not aim at the detailed annotation of gestures and mimics. Head movements, hand and arm movements, as well as shoulder movements were considered gestures; lip and eyebrow movements were noted as mimics. Other movements did not occur in the corpus, because of the seating position of the participants during the reading task. Eye movements could not be considered, as our study supposed reading texts printed on paper. No eye tracking technique was applied. This type of annotation is not detailed, but it allows making one's own analysis of the concrete parameter(s) applying deeper annotation techniques if needed.

Such analysis was performed in a series of pilot experiments conducted to compare our data with the results obtained on the material of other languages [27, 28]. The results of the experiments demonstrated that synchronization of mimics and gestures with the intonation centre (focus) may be different in laboratory speech and in actor's speech. It is often shifted in relation to the focus in laboratory speech, while synchronized with the nucleus in actors' speech. In both types of speech, the same paralinguistic phenomena may occur, but their frequency differs.

Evaluation of Irony by Native Listeners. A series of perceptual experiments was conducted in order to validate the methodology of the corpus construction [29–31]. In these experiments ironic and non-ironic target fragments were presented to the listeners with no indication of ironic or non-ironic meaning, i.e. they were given without any lexical marker or context. The same methodology may be found in other works as well [37]. Forty-five fragments were selected for each experiment and then randomized. The task suggested to the listeners was to associate the fragment with one of the written texts or dialogues with ironic or non-ironic meaning. Thus, the participants did not have any idea about the experimental task. To avoid possible mistakes due to the misunderstanding of the term "irony", this term was not used (as it could be confused, on the one hand, with humour and on the other hand, with doubt).

The results showed that the texts and dialogues elaborated as reading tasks provided good communicative paradigms, which enabled the irony expression well understood by the native listeners. In the target fragments that were evaluated by listeners the percentage of recognition was annotated.

4.2 Information Included in the File Name

Some information is indicated in the file name:

- Speaker gender;
- Ironic or non-ironic meaning;
- Sentence type;
- The presence of a marker (such as "tozhe mne"—"some… you are") or a remark (she sad mocking) in the preceding or following contexts, if any.

5 Preliminary Results

5.1 Prosodic Features of Irony

At the moment, we have conducted acoustic analysis and analysis through synthesis and modification of the melodic contour of the target fragments. The acoustic analysis of the ironic utterances from the corpus showed that there are three most salient characteristics of irony expression, which are also relevant for its perception [29–31]. In most cases (72%), the increase in the intensity level differerentiated ironic fragments from the non-ironic ones. Less often (25%) there was a decrease in the same parameter. Second phonetic property that differentiated ironic utterances was the contrast in the stressed vowel duration. Most often, the stressed vowel was elongated (81%); it could also be shortened, although these cases occurred less frequently in the material (10%). 46% of ironic utterances had a notably enlarged melodic range and 35% had a reduced melodic range. The experiments, in which various modifications of the speech signal (notably of its pitch contour) were performed, showed that the pitch pattern plays an important role in irony production and perception. However, this parameter is not the only one that helps to express the ironic meaning.

5.2 Intonation Patterns

We compared the distributions of melodic types in non-ironic (neutral) and ironic contexts, across various sentence types. It should be noted that the contexts termed "neutral" often sounded rather expressive.

Narratives. In standard Russian, narratives (declarative sentences) are most often produced with falling intonation. In our material the following statistics were observed. In neutral contexts, simple falls with varying falling intervals occurred in 52% cases. Other frequent melodic types are the emphatic fall (model 02), 24%, and the low (fall)-rise, 9% (model 13). The latter two correspond to expressive realizations of the narratives. In ironic contexts, there were 35% of falling tones. Similar to neutral contexts, the emphatic fall (model 02) was produced in 22% cases. The specific feature of ironic contexts is high frequency (30%) of the model 04: falling intonation with a wider interval of the falling tone, a higher level of intensity and the corresponding voice quality (used in exclamations). Additionally, the low (fall)-rise is very rare in ironic statements.

Interrogatives. In both types of contexts, the most frequent model is the standard rise-fall typical for yes-no questions (model 07): 86% in neutral contexts and 60% in ironic contexts. A variant of this rise-fall, with the melodic peak shifted to the right (model 07b), occurred with almost equal frequency: 11% in neutral contexts and 10% in ironic contexts. It should be noted here that this model is described as expressive and may carry various connotations [32, 34]; however, our results show that these connotations may be ironic and non-ironic. In terms of melodic types only, ironic and neutral utterances may not differ at all.

Figures 1 and 2 present a minimal pair: the utterance "Malen'kaya?" ("Small?") produced in neutral and ironic context by the same speaker. In both cases we observe a rise-fall with peak located to the right of the stressed vowel (model 07b).

Exclamations. The most frequent model for exclamations is model 04: falling intonation with a wider interval of the falling tone, a higher level of intensity and the corresponding voice quality—the model that is also observed in ironic narratives (see above). The frequency of model 04 is 46% in neutral contexts and 22% in ironic contexts. The emphatic fall occurs in 13% and 9%, respectively. Ironic contexts differ in the frequency of simple falls (models 01, 01a, 01b): 19% in neutral contexts versus

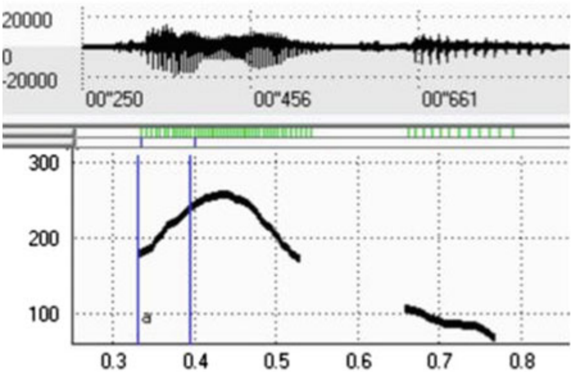


Fig. 1 Example of the interrogative non-ironic utterance "Malen'kaya?" (Little?); male speaker; Wave Assistant software

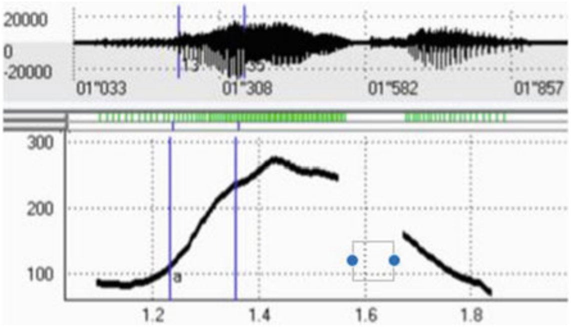


Fig. 2 Example of the interrogative utterance with irony "Malen'kaya?" (Little?); male speaker; Wave Assistant software

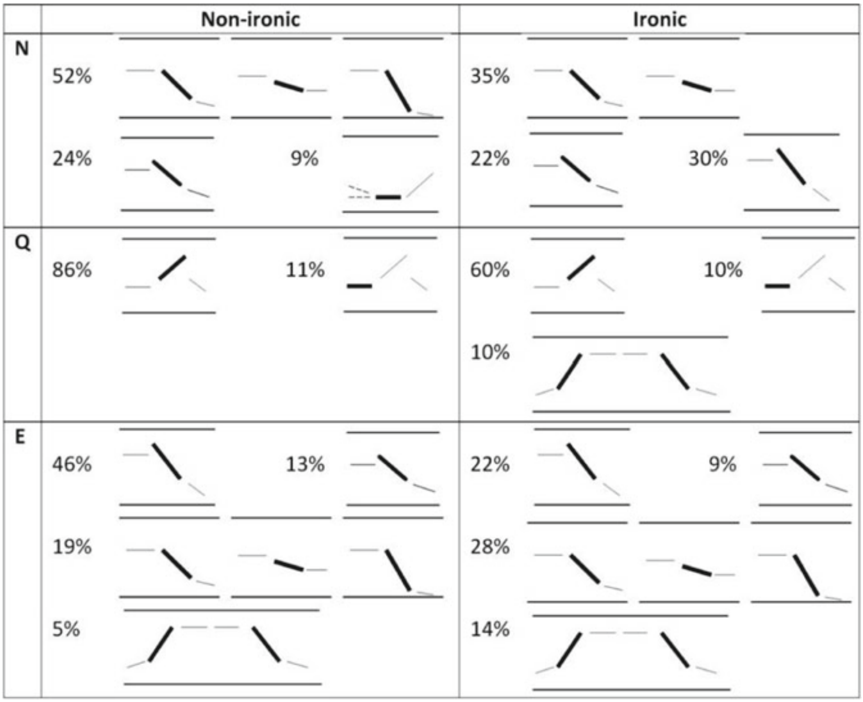
28% in ironic contexts. Being often accompanied by greater duration, low fall is one of the typical ways to pronounce an ironic exclamation.

Another difference is the frequency of the model 05: a pitch rise to a very high level at some point in the pre-nuclear part, high plateau sustained up to the nucleus, a fall on the nucleus, normally accompanied by high intensity. This model is observed in 5% cases in neutral contexts and in 14% cases in ironic contexts.

Some of the target utterances were placed in contexts with adjacent ironic markers. In our material, the melodic pattern produced within the target utterance matched that of the ironic marker in 24% cases. However, in those contexts the speakers used a very limited number of melodic patterns: only five. That is, the value of 24% is just above chance, which does not allow us to confirm the initial hypothesis.

Summing up the results of the analysis, we may assume that there are some preferences in speakers' choice of the melodic type when pronouncing an utterance with irony (see Table 2). Simultaneously, to express irony, speakers may use any kind of deviation from the neutral realizations. One of the possible ways is to use a melodic pattern typical for another sentence type: e.g., pronouncing a question or a narrative as an exclamation, and exclamation—as a statement (with a very low fall).

Table 2 Most frequent melodic patterns in non-ironic and ironic utterances in the corpus, grouped in terms of sentence type: narrative (N), question (Q), exclamation (E)



	Non-ironic	Ironic
N	52% 24% 9%	35% 22% 30%
Q	86% 11%	60% 10% 10%
E	46% 13% 19% 5%	22% 9% 28% 14%

5.3 *Spectral Density*

The power spectral density was higher in ironic utterances than in the homonymous non-ironic utterances (the direction of pitch movement was taken into account), as it can be seen from Fig. 3, which represents the comparison of the total number of ironic and non-ironic utterances. The formant components prior to 4 kHz were reinforced. The difference observed between sentence types was the F0 peak shift in exclamatory sentences (Fig. 6), while in other sentence types the formants (F1, F2, F3) had more energy in ironic utterances in comparison with the non-ironic ones without such shifts (Figs. 4, 5 and 6).

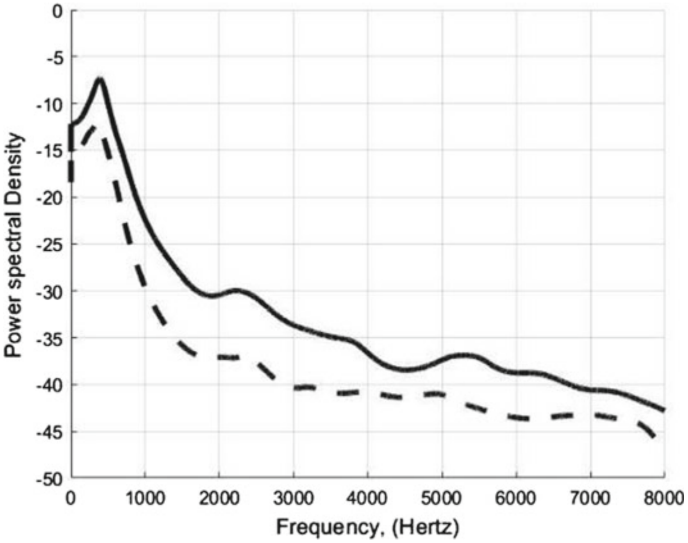


Fig. 3 Spectral density in ironic (solid line) and non-ironic (dotted line) sentences

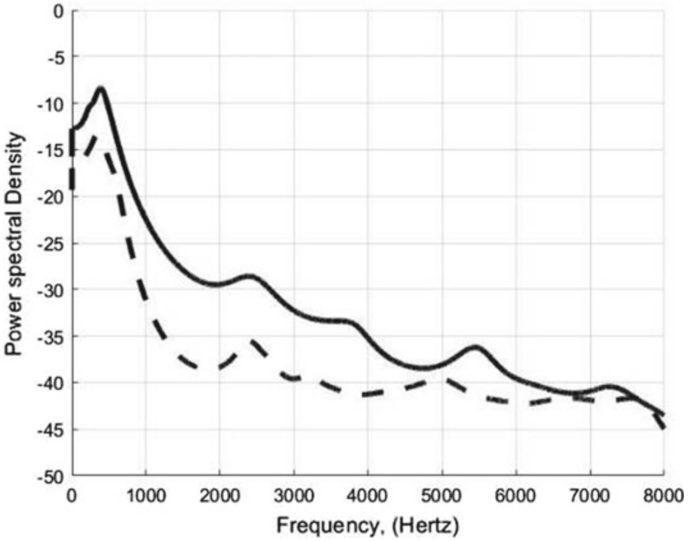


Fig. 4 Spectral density in narrative ironic (solid line) and non-ironic (dotted line) sentences

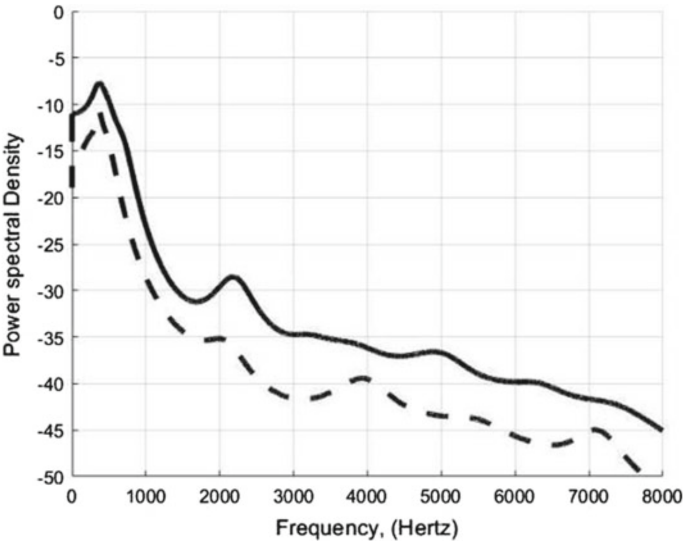


Fig. 5 Spectral density in interrogative ironic (solid line) and non-ironic (dotted line) sentences

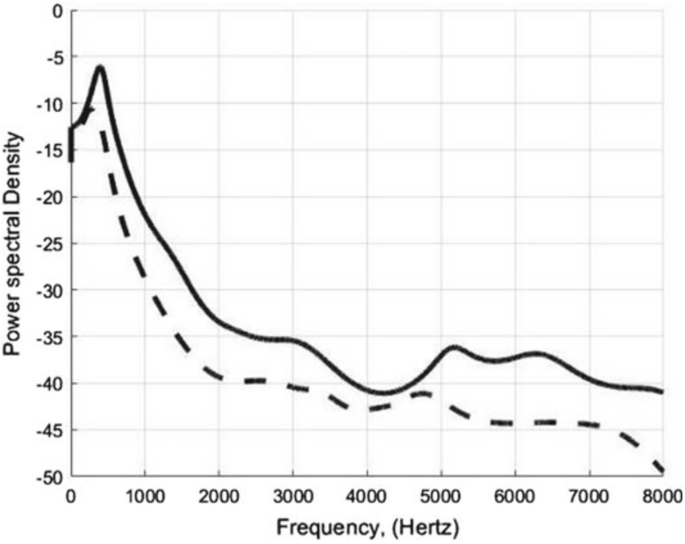


Fig. 6 Spectral density in exclamatory ironic (solid line) and non-ironic (dotted line) sentences

6 Conclusion

The obtained multimedia corpus of Russian ironic speech allows both basic and applied research. On the one hand, it enables studies of the pronunciation standard in Russian, notably its intonation norm, as well as intra- and interspeaker variability. On the other hand, it may help to construct probabilistic models of acoustic cues of irony.

The part of the material containing evaluation by native listeners may be very interesting to and useful for the developers of the automatic systems of speech recognition. Any tool, including neural networks, requires a control set of data for the validation of its effectiveness. In our corpus, such a set, approved by Russian native speakers, already exists. It allows the usage of corpus not only by linguists, but also by specialists developing human–computer dialogue systems.

Thus, we can conclude that the obtained corpus may be used for various purposes and is valuable not only for phonetic and linguistic research but also for cognitive studies, psychology, speech technologies and e-communication field.

Acknowledgements This research was supported by the RFBR grant No. 20-012-00552.

References

1. Cutler, A. (1974). On saying what you mean without meaning what you say. In *Proceedings from the 10th Regional Meeting of the Chicago Linguistic Society* (pp. 117–123). Chicago: CLS.
2. Giora, R. (1995). On irony and negation. *Discourse Processes, 19*(2), 239–264.
3. Haverkate, H. (1990). A speech act analysis of irony. *Journal of Pragmatics, 14*, 77–109.
4. Ivanko, S. L., & Pexman, P. M. (2003). Context incongruity and irony processing. *Discourse Processes, 35*(3), 241–279.
5. Joshi, A., Sharma, V., & Bhattacharyya, P. (2015). Harnessing context incongruity for sarcasm detection. In *Proceedings of 53rd Annual Meeting of the Association for Computational Linguistics*, Beijing, China (p. 757).
6. Maynard, D., & Greenwood, M. A. (2014). Who cares about sarcastic tweets? Investigating the impact of sarcasm on sentiment analysis. In *Proceedings of LREC*.
7. Barbieri, F., Saggion, H., & Ronzano, F. (2014). Modelling sarcasm in Twitter, a novel approach. In *2014, ACL* (p. 50).
8. Carvalho, P., Sarmento, L., Silva, M. J., & De Oliveira, E. (2009). Clues for detecting irony in user-generated contents: Oh...!! it's so easy;-). In *Proceedings of the 1st International CIKM Workshop on Topic-Sentiment Analysis for Mass Opinion* (pp. 53–56). ACM.
9. Davidov, D., Tsur, O., & Rappoport, A. (2010). Semi-supervised recognition of sarcastic sentences in Twitter and Amazon. In *Proceedings of the Fourteenth Conference on Computational Natural Language Learning* (pp. 107–116). Association for Computational Linguistics.
10. González-Ibánez, R., Muresan, S., & Wacholder, N. (2011). Identifying sarcasm in Twitter: A closer look. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: Short Papers* (Vol. 2, pp. 581–586). Association for Computational Linguistics.
11. Liebrecht, C., Kunneman, F., & van den Bosch, A. (2013). The perfect solution for detecting sarcasm in tweets # not. In *2013, WASSA* (p. 29).

12. Camblin, C. C., Gordon, P. C., Swaab, T. Y., et al. (2007). The interplay of discourse congruence and lexical association during sentence processing: Evidence from ERPs and eye tracking. *Journal of Memory and Language, 56*(1), 103–128.
13. Mishra, A., Bhattacharyya, P., & Kanojia, D. (2016). Predicting readers' sarcasm understandability by modeling gaze behavior. In *Cognitively inspired natural language processing* (pp. 99–115). Springer Verlag.
14. Filik, R., Leuthold, H., Wallington, K., & Page, J. (2014). Testing theories of irony processing using eye-tracking and ERPS. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 40*(3), 811–828.
15. Michael, St., & Zahra, A. (2019). Automatic sarcasm detection with textual and acoustic data. *International Journal of Recent Technology and Engineering, 8*(4), 1357–1360.
16. Bryant, G., & Fox Tree, J. (2008). Is there an ironic tone of voice? *Language and Speech, 48*, 257–277.
17. Niebuhr, O. (2016). Rich reduction: Sound-segment residuals and the encoding of communicative functions along the hypo-hyper scale. In *7th Tutorial and Research Workshop on Experimental Linguistics* (pp. 11–24). St. Petersburg, Russia.
18. Cheang, H., & Pell, M. (2008). The sound of sarcasm. *Speech Communication, 50*(5), 366–381.
19. Cheang, H., & Pell, M. (2009). Acoustic markers of sarcasm in Cantonese and English. *The Journal of the Acoustical Society of America, 126*(3), 1394–1405.
20. Del Ré, A., Hirsch, F., & Dodane, Ch. (2018). L'ironie dans le discours: des premières productions enfantines aux productions des adultes. *Cahiers de Praxématique, 70.*
21. Scharrer, L., & Cristmann, U. (2011). Voice modulations in German ironic speech. *Language and Speech, 54*(4), 435–465.
22. Schaffer, R. (1982). *Vocal clues for irony in English.* Ph.D. thesis, Ohio State University.
23. Kodzasov, S. (2009). *Studies on the prosody in Russian [Issledovaniia v oblasti russkoi prosodii].* Iazyki slavianskikh kultur.
24. Wagner, P., Malisz, S., & Kopp, S. (2014). Gesture and speech in interaction: An overview. *Speech Communication, 57*, 209–232.
25. Barbulescu, A., Ronfar, R., & Bailly, G. (2017). Generative audio-visual prosodic model for virtual actors. In *EEE engineering in medicine and biology magazine: The quarterly magazine of the Engineering in Medicine & Biology Society* (pp. 40–51).
26. Bergmann, K., Aksu, V., & Kopp, S. (2011). The relation of speech and gestures: Temporal synchrony follows semantic synchrony. In *Proceedings of GESPIN 2011: Gesture and Speech in Interaction*, Bielefeld, Germany.
27. Chui, K. (2005). Temporal patterning of speech and iconic gestures in conversational discourse. *Journal of Pragmatics, 37*, 871–887.
28. Loehr, D. (2012). Temporal, structural, and pragmatic synchrony between intonation and gesture. *Journal of the Association for Laboratory Phonology, 3*, 71–89.
29. Skrelin, P., Kochetkova, U., Evdokimova, V., & Novoselova, D. (2020). Can we detect irony in speech using phonetic characteristics only?—Looking for a methodology of analysis. In *Proceedings of the 22nd International Conference SPECOM, LNAI* (Vol. 12335, pp. 544–553). Heidelberg: Springer.
30. Kochetkova, U., Skrelin, P., Evdokimova, V., & Novoselova, D. (2020). Perception of irony in speech. In O. V. Sherbakova (Ed.), *Proceedings of the 4th International Conference on Neurobiology of Speech and Language* (pp. 72–73). Saint Petersburg: Skifia-Print.
31. Kochetkova, U., Skrelin, P., Evdokimova, V., & Novoselova, D. (2021). The speech corpus for studying phonetic properties of irony. In *Language, music and gesture: Informational crossroads. LMGIC 2021* (pp. 203–214).
32. Volskaya, N., & Kachkovskaia, T. (2016). Prosodic annotation in the new corpus of Russian spontaneous speech CoRuSS. In *Proceedings of the Speech Prosody 2016* (pp. 917–921).
33. Bryzgunova, E. A. (1977). *Sounds and intonation of Russian speech [Zvuki i intonacija russkoj rechi].* Russkij jazyk. (in Russian).
34. Kachkovskaia, T., Kocharov, D., Skrelin, P. A., & Volskaya, N. B. (2016). CoRuSS—A new prosodically annotated corpus of Russian spontaneous speech. In *Proceedings of LREC 2016* (pp. 1949–1954).

35. Brugman, H., Wittenburg, P., Levinson, S. C., & Kita, S. (2002). Multimodal annotations in gesture and sign language studies. In *Third International Conference on Language Resources and Evaluation* (pp. 176–182).
36. Kipp, M. (2009). Multimodal annotation, querying and analysis in ANVIL. In *Multimedia information extraction* (pp. 351–368). Wiley.
37. Nauke, A., & Braun, A. (2011). The production and perception of irony in short context-free utterances. In *Proceedings from the 17th International Congress of Phonetic Sciences (ICPhS)* (pp. 1450–1453). Hong Kong, China: ICPhS.

Theory of Mind and the Mechanism of Imagination for a Companion Robot

Artemiy Kotov and Anna Zinina

Abstract We extend the architecture of F-2 companion robot to simulate some verbal (semantic) judgements, that can be considered as the theory of mind—other's thoughts, or as a verbal imagination—unreal emotionally significant situations, that could be constructed by the robot. F-2 has a syntactic parser and text comprehension engine, based on productions, where each incoming sentence meaning or a computer vision event is associated with the most relevant script. The robot assigns to the theory of mind scripts with low relevance, where the valency of experiencer is occupied by the opponent: "he may experience that". This script is not relevant for the subject, but could represent the emotional experience of the addressee. For the imagination, we consider irrelevant scripts, that generate an unreal but emotionally significant situation for a given fact. In both cases for a given stimulus the robot activates a number of scripts and selects the scripts with low relevance to the stimulus, but with high emotional significance.

Keywords Model of consciousness · Text understanding · Text semantics · Emotional robot companions · Cognitive processing

1 Introduction

The imagination and the theory of other's thoughts—Theory of Mind [1], are the promising components for a companion robot. In this work we suggest a solution to simulate some parts of the verbal imagination, as well as some parts of the verbal theory of mind (TOM). For a given situation, the theory of mind may suggest *what another person could think?* and the imagination can construct an unreal emotional extension or modification to the current situation, like *what could*

A. Kotov (✉) · A. Zinina
Kurchatov Institute, Moscow 123182, Russia
e-mail: kotov@harpia.ru

Russian State University for the Humanities, Moscow 125047, Russia

Moscow State Linguistic University, Moscow 119034, Russia

P. Eismont et al. (eds.), *Literature, Language and Computing*,
https://doi.org/10.1007/978-981-99-3604-5_20

happen but did not? In these studies, we rely on the software architecture for text and event processing for the companion robot F-2 (Fig. 1). This experimental robot is designed with extended software architecture, that includes linguistic processor for natural text (supporting communication in Russian) and an inference engine, that simulates emotional dynamics, suggests verbal responses and behavioral patterns, to be performed by the robot in communication. The inference engine may receive semantic representations from written and oral texts as well as representations for surrounding events, recognized by computer vision modules.

The inference engine of F-2 robot relies on the classic architecture of productions (scripts). The idea to use productions as the basic emotional processor for robots had been suggested by M. Minsky as the mechanism of *proto-specialists* [2] and further extended by A. Sloman in CogAff architecture [3]. Within this approach, each incoming event is evaluated in order to select the most relevant script to process it. Although Sloman argues, that several scripts from different domains (e.g. emotional and rational scripts) can operate in parallel, in many applied cognitive architectures the alternative scripts are usually discarded: an event is processed by the most relevant script, which suggests the output reaction for an agent or robot.

Within the F-2 architecture for each incoming fact—utterance semantics or computer vision event—the robot calculates distances to the existing semantic representations: script premises. While the first script—closest to the stimulus—is naturally considered as the most relevant, we suggest the approach, where other scripts are not discarded. If some minor script assigns a valency of *experiencer* to some other person in communication (e.g. the addressee) and adds an emotional evaluation, the robot may consider this as a possible view by that person. On the other side, some irrelevant scripts offer an emotional view on the situation from the position of the robot: the robot may believe, it is at the center of public attention (*Everyone is paying attention at me!*) or that the situation is an encounter where the robot can beat the opponent (*I shall win!*). These irrelevant representations are evaluated lower, than the first script—"true" representation of events to the robot. However, we suggest

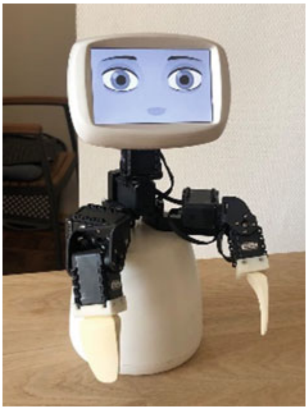


Fig. 1 F-2 robot

to assign them to some other domains like 'other's thoughts' (TOM) or 'imaginary world'.

2 Linguistic Processing

We are developing a multi-component parser to analyze the natural language text for the F-2 robot. The components—*morphological*, *syntactic*, and *semantic*—correspond to the levels of a classic language model. The robot receives a text in written form, as a file or via the RSS subscription (e.g. for daily news processing), or in an oral form in direct communication, in this case an external speech-to-text service is used. In case, several variants are returned by the speech-to-text processor, they are analyzed in parallel and the best variant is used further for script analysis. The text is analyzed sentence by sentence.

2.1 Morphological Processing

At the stage of morphological analysis, all words in a sentence are lemmatized (reduced to the initial form), and each word is tagged with grammemes (for the categories *case*, *number*, *tense*, etc.) All the possible cases of homonymy are constructed to be resolved at subsequent stages of analysis. The system of morphological analysis includes three components: (1) a dictionary of word forms in SQL that contains information about 98 thousand lexemes (based on the OpenCorpora resource dictionary [4]), (2) a system for analyzing numbers and alphanumeric complexes (e.g. *19-th*, *25-year-old*), which is based on 40 regular expressions, and (3) a neural network algorithm (*guesser*) for predicting morphological features of unknown words. These three components work sequentially: a wordform is passed to the subsequent component if no parsing result has been suggested by the previous one. Consider the sentence *Linguists have tormented psychologists at the conference with stupid questions* (sentence and the results of analysis are translated from an original language). The wordform *psychologists* [*psihologi*] is reduced to its initial form *psychologist*, and being a homonym, is tagged with two sets of grammemes that differ in grammatical case: ***accusative/genitive***, *noun, animate, masculine, plural.*

2.2 Syntactic Analysis

At the syntactic stage, the words are sequentially added to a parsing stack, after which syntactic rules are applied to the top of the stack. The goal is to "collapse" all the words of the sentence into a connected syntactic tree. We rely on the syntactic model by Gladky [5], which combines the principles of dependency grammar and phrase structure grammar. The parser uses 620 grammar rules in the SyntXML format [6]. Each rule describes a sequence of segments to be bound by a syntactic link. A rule may check grammatical features, grammatical agreement of neighboring segments, and so on. After a rule is successfully applied, all its segments are bound by a syntactic link where the head is either one of these segments or a *zero* head, see [5]. All grammar rules are applied sequentially to the same stack, and even if one rule is successful, the application of rules continues, that allows the system to handle the cases of syntactic ambiguity. Morphological and syntactic ambiguity is accounted by the duplication of stacks, so the parser operates on a set of stacks at the same time: usually up to 512 stacks for a sentence. On the next stage this would allow to recognize the required meaning (script) in any of the ambiguous meanings a sentence, for example, for perspective handling of computational humor. The result of the parsing stage is a set of syntactic representations of a particular sentence. Figure 2 represents a syntactic tree of the sentence *Linguists have tormented psychologists at the conference with stupid questions*. The head of the tree is the predicate *to torment* and all other lexemes are subordinate to the head. At this stage of analysis, the *accusative* variant of *psychologists* was chosen rather than the *genitive* because the predicate *to torment* [*zamuchit'*] is not able to subordinate a genitive noun. At the same time this transitive verb can be bound with an accusative noun. Thus, some morphological homonymy that was not resolved by the morphological component, can be resolved at the syntactic stage with the help of grammar rules.

2.3 Semantic Analysis

At the semantic stage, the parser builds a semantic representation for each syntactic tree. The semantic representation of a sentence consists of predications (frames), which are added to the knowledge base. When constructing a semantic representation, the results of the two previous stages of linguistic analysis are considered. At the stage of morphological analysis, semantic features are assigned to the words of the sentence along with grammatical features. Semantic features are also extracted from the dictionary of wordforms. We use a list of 927 semantic features developed on the basis of the Russian semantic dictionary [7]—among them are ‘person’, ‘to-touch’, ‘be-expensive’, etc. We also use a list of 3900 semantic features developed automatically with the help of the word embedding model trained on the Russian-language Wikipedia and The Russian National Corpus [8]. Using this model, similar words (the words whose vectors are close to each other in the vector space) were grouped and received a common marker. These markers have names starting with *like-* or the @ symbol (see Table 1). 37 thousand lexemes in the dictionary were annotated by semantic attributes. At the stage of syntactic analysis, semantic valencies are assigned to segments in the syntactic tree. We use a list of 23 semantic valencies, based on [9]. Among the valencies are *predicate*, *agent*, *patient*, *instrument*, etc. Thus, at the stage of semantic analysis the sentence is divided by valencies, and each valency receives the semantic markers of the corresponding words. For a compound

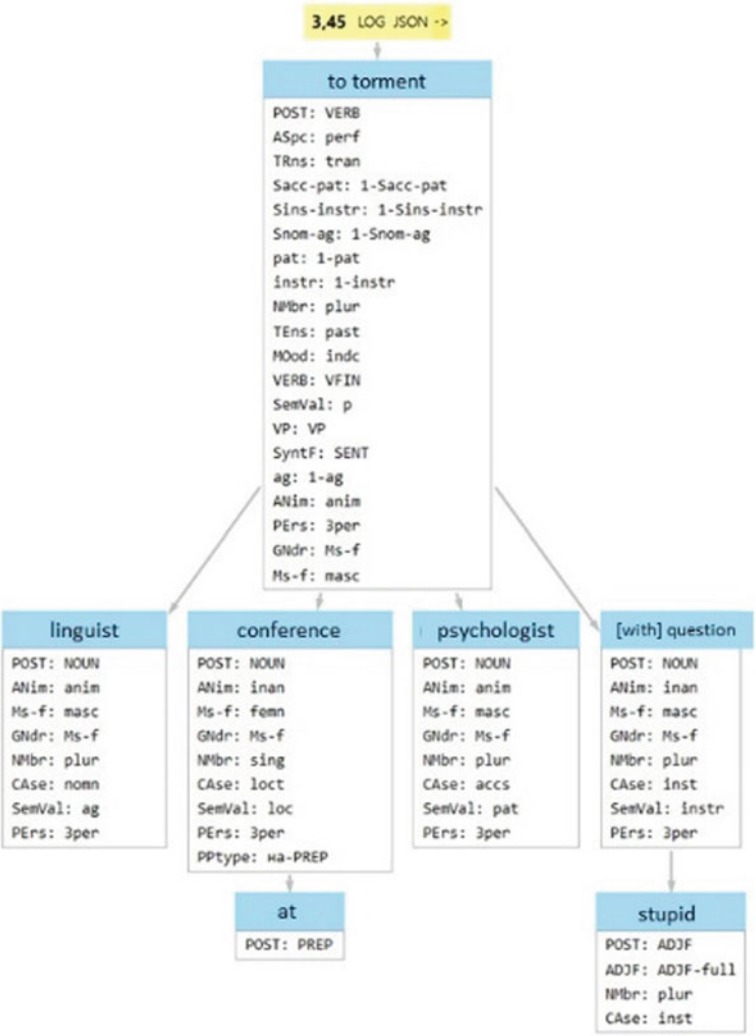


Fig. 2 Syntactic tree: *Linguists have tormented psychologists at the conference with stupid questions* (lexemes here are translated)

sentence, several frames are constructed. These semantic representations are easily stored in a database, and available for search queries, for example, for the task of question answering for the robot.

Table 1 represents a semantic frame for the sentence *Linguists have tormented psychologists at the conference with stupid questions.* Each column of the table is a valency filled with semantic markers. This table is a representation of the result of linguistic analysis of the incoming sentence. Using data like this, the F-2 robot can

Table 1 Semantic representation (frame) for the sentence *Linguists have tormented psychologists at the conference with stupid questions*. Note, that the word *question* has two polysemic meanings: 1 and 2

p: *to torment*	**ag:** *linguist*	**pat:** *psychologist*	**instr:** *question*	**loc:** *conference*
1 past tense 1 assertive 1 cause harm 1 like execute 1 @89_ VERB	1 plural 1 someone 1 profession 1 profession-in-science 1 like philologist 1 @441_NOUN	1 plural 1 someone 1 profession 1 like psychologist 1 @200_NOUN	1 plural 1 abstract object 1 neg. situation 1 frustration 1 like issue 1 @364_NOUN 2 plural 2 abstract object 2 message 2 request 2 like request 2 @364_NOUN	1 singular 1 container 1 abstract container 1 abstract object 1 organized time 1 like symposium 1 @86_NOUN 1 at_preposition

"understand" the situation and, depending on its settings and previous communication experience, associate itself with the annoying agent, the tortured patient, or an independent observer, that in turn would determine robot's responses and reactions to this fact.

3 Scripts Engine

Visual events and semantic frames from texts are processed via an inference engine, which contains a list of scripts: *if-then* operators or *productions*. As suggested within CogAff architecture [3], scripts are divided into groups for emotional and rational processing. *D-scripts* (or *dominant scripts,* n = 72) are responsible for emotional processing on the primary reactive level, and *r-scripts* (or *rational scripts*, n = 3500) are used for solving the ambiguity and for rational processing of incoming events and inferences—on some upper deliberative level. In this respect, d-scripts correspond to the notion of *proto-specialists* by Minsky [2] and *r-scripts* refer to the original concept of scripts by Schank [10]. In addition to that, rule-scripts describe the rules of behavior of the robot in different situations, not directly connected to the emotional arousal, like saying *sorry* or trying to cheer up a depressed addressee.

For each script, its *if*-condition (premise) and *then*-condition (inference) are represented as a set of semantic markers, divided into valencies. This representation corresponds to the semantic frame (as in Table 1), but is always monosemantic and does not include the indicators of ambiguity *sem/polysem.* For each incoming semantic frame (stimulus) the system calculates its similarity to all the premises of the existing scripts. For an ambiguous word in speech, all the sem/polisem values are taken into account—for each script the closest semantics will be selected in case of ambiguity. For each script, the number of its satisfied valencies is calculated, and a modified

Jaccard index is used to evaluate the degree of similarity for each valency. Let A be the number of semantic markers in a valency of a script premise, and C—the set of markers, present both in the stimulus and in the premise for the selected *sem/polysem*. Then the similarity of a pair of valencies is calculated as:

$$v = \max_{sem/polysem} \frac{\sum_C m}{\sum_{A\backslash C} a + \sum_C n}\Bigg|_{sem/polysem} \tag{1}$$

where v—in the degree of similarity, *sem/polysem* is the attribute of the stimulus, indicating an ambiguous meaning for incoming speech semantics; m—the degree of similarity of semantic markers from C, a—is a normalizing coefficient for the markers from $A\backslash C$; n—normalizing coefficient for the markers from C.

Following the calculated similarity between the stimulus and the premises, the scripts are sorted depending on the reduction of their relevance: most congruent and relevant scripts come first (see Fig. 3). For the simulation of behavior, the relevant d-scripts are activated proportionally to its congruency to the stimulus and proportionally to its sensitivity. An activated script sends to an output the attached behavioral packages in BML format [11]. The robot can execute BML packages for the scripts with the highest activation. A script is released as soon as its BML packages are executed on the robot. Scripts with minor activation can also send their BML packages to the robot, they are executed when the robot does not have any major stimulus to process, like in a situation of inactivity. Scripts also reduce their activation in time, so minor scripts can be discarded even if their BML packages are not executed on the robot. At the same time, the combination of major and minor scripts may result in emotional blending on the robot [12], when, for example, a response to a user's question and user's gaze is performed by head/eyes and mouth of the robot, while internal tension is expressed with hands in the form of scratching and manipulation with hands.

D-scripts were developed as emotional semantic invariants in mass media texts, texts of election campaigns and advertising, collected since 1999 [13]. In theoretical

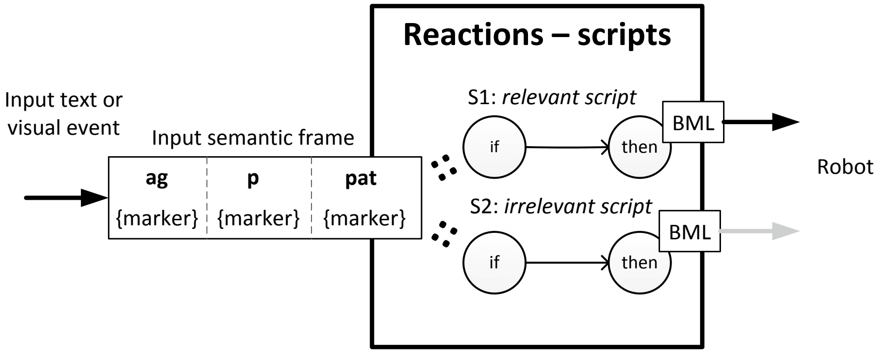


Fig. 3 Activation of scripts by the input semantic frames, sorting of scripts following their relevance

analysis d-scripts are defined similar to frames in FrameNet project [14]. Within the computer system of F-2 robot d-scripts are also defined as sets of semantic markers (as in Table 1). Negative scripts contain main valencies AGGR—an aggressor in a negative situation, and VICT—a victim. The experiencing subject has to associate himself with VICT, who is the experiencer in a negative situation. For example, PLAN d-script is defined as 'AGGR is making some evil plans against VICT' and SUBJV d-script as 'VICT worries, that AGGR thinks only about one thing' (*You think only about your football!*). Depending on the distribution of valencies, negative d-scripts can be used in a conflict—*You always think about yourself!* for negative speech influence—*Government thinks only about themselves!* etc. The definitions of positive d-scripts include the valencies EXP—experiencer, and BON—protagonist. For example, ATTENTION d-script is defined as 'EXP is happy, that BON pays attention at him'. Positive d-scripts can be used to express own emotions—*Great, that you have noticed my new bag!*, for a compliment—*Everyone has noticed your new bag!* or for advertising—*You should buy this bag, everyone is looking at!* At the same time, an incoming phrase *John is looking at you!* can activate all the indicated negative and positive scripts: the subject may be happy to be at the center of attention and worry, that *John* 'makes some evil plans' or 'does not look around'. So, the sensitivity of scripts may be used to tune the robot's emotional profile: create an optimistic robot, preferring positive d-scripts or a depressive robot, which prefers the negative d-scripts. In this respect, the utterance *Linguists have tormented psychologists* is rather controversial: it may provoke negative or positive emotional perception as well as different etiquette replies, defined by rule-scripts (see Table 2).

In fact, the robot does not have any pre-disposition to associate itself with linguists or psychologists. As seen in the table, it suggests both (a) etiquette replies (*I'm sorry!*) and emotional replied (*We'll torture everyone!*) for linguists as well as (b) etiquette replies (*You* [linguists] *did nothing bad!*) and emotional replies (*You are killing me!*) for psychologists. Here the robot gets into an ambiguous situation, where the protagonist is not defined: who is the robot—a linguist or a psychologist? We consider that scripts in Table 2 are evaluated depending on their **relevance to the situation**, not relevance to protagonist. If the protagonist is defined (e.g. *Linguists have troubled robot!*) its semantic markers are used in evaluation and the scripts with correct valency distribution are preferred: "linguists have troubled me, but I have to say *I'm even interested in it!* or *You* [linguists] *did nothing bad!*" As we propose, the evaluation of scripts, relevant to the situation, may still be used to select the semantic representations for the Theory of Mind and for imagination.

3.1 Model for the Theory of Mind

Theory of mind represents the judgements or emotional evaluations, assigned to another participant of the situation. For the theory of mind, we shall try to simulate emotional evaluations of the situation, that could be suggested by d-scripts. D-scripts

Table 2 Acitivated scripts and reactions to an incoming atterance *Linguists have tormented psychologists*

Similarity	Script	Script type	Identification with valency	Speech output
0.1676	ACCEPTANCE: to represent the situation as positive to me	Rule	'Psychologist'	*I'm even interested in it!*
0.1643	RESOLVING THE PROBLEM (my fault): to offer compensation	Rule	'Linguist'	*Let me buy you* [the psychologist] *a cake for this!*
0.1643	CONCILIATION (my fault): explicit acceptance of an incorrect action	Rule	'Linguist'	*I probably have tortured you* [the psychologist]
0.1603	CONCILIATION (his/her fault): reduce the categorical nature of the situation	Rule	'Psychologist'	*You* [the linguist] *did nothing bad!*
0.1545	PLANNING	Negative d-script	'Psychologist', AGGR	*You* [the linguist] *want to drive me crazy!*
0.1524	PROTECTION: if the protagonist is in danger	Positive d-script	'Psychologist—protagonist', <empty valency>	*Don't bother him* [the psychologist]! (to the 'linguist')
0.1511	ETIQUETTE (my fault)	Rule	'Linguist'	*I'm sorry!*
0.1511	CONCILIATION (my fault): reduce the categorical nature of the situation	Rule	'Linguist'	*I didn't mean to!*
0.1511	WE•INADEQ: worry about my own inadequacy	Negative d-script	'Linguist', AGGR/VICT	*I'm doing something wrong!*
0.1495	WE•DANGER: happy to conquer others	Positive d-script	'Linguist', BON/EXP	*We'll torture everyone!*
0.1482	RULE (my fault): formulate a rule for myself	Rule	'Linguist'	*I shouldn't be so annoying*
0.1414	RULE (his/her fault): teach another person	Rule	'Psychologist'	*You* [the linguist] *shouldn't be so annoying*

(continued)

Table 2 (continued)

Similarity	Script	Script type	Identification with valency	Speech output
0.1243	EMPATHY: to sympathize with the protagonist	Positive d-script	'Psychologist–protagonist', <empty valency>	*Poor* [psychologists]*!*
0.1171	DANGER: he makes me bad—reply with aggression or flee	Negative d-script	'Psychologist', VICT	*You* [the linguist] *are killing me! I'll get my revenge on you!*

assume, that the participant in communication assigns himself to the emotionally relevant valency: EXP or VICT. Consider the scripts, where the protagonist is defined and is assigned to EXP or VICT valency. These are the relevant scripts: the agent evaluates the situation from his own point of view, or the point of view of the corresponding protagonist. If for some "neutral" input *Linguists have asked psychologists* the robot associates itself with 'psychologists', it may invoke positive scripts (*I'm happy to be asked!*—ATTENTION) or negative scripts (*I'm troubled to be asked!*—DANGER), see Fig. 4. These are the scripts with congruent distribution of valencies: protagonist = EXP or VICT. Let us consider the scripts with incongruent distribution of valencies: me/protagonist does not correspond to EXP/VICT. In the suggested model we allocate these scripts to form the theory of mind of the opponent. So, if the robot is a 'psychologist', it may construct a judgement, assigned to the 'linguist': *the linguist can be happy, as he has helped me* (ASSIST) or *he can be troubled, as he thinks, I'm inadequate* (INADEQ). Such scripts, relevant to the situation, but with incongruent distribution of valencies may be assigned to the opponent, co-referential with EXP or VICT.

3.2 Model of Imagination

We suggest that emotionally significant scripts may be allocated to imagination, even if they are not relevant to the situation: it means, that for a given input the robot has constructed some other, higher evaluated scripts. Consider that *a linguist asks the psychologist*: this situation can be evaluated by an r-script ('a person asks a person'), that would be a "true" representation of the event for the robot. At the same time, with some lower relevance it can be associated to a d-script, e.g. ATTENTION, which can suggest a situation 'the linguist compliments me/psychologist'. If the distance between the two scripts is rather high, they are incompatible: the latter script cannot suggest a "true" representation of the event: the linguist does not compliment the psychologist (he only asks a question). However, if this false representation is emotionally relevant, it is allocated to the mechanism of imagination,

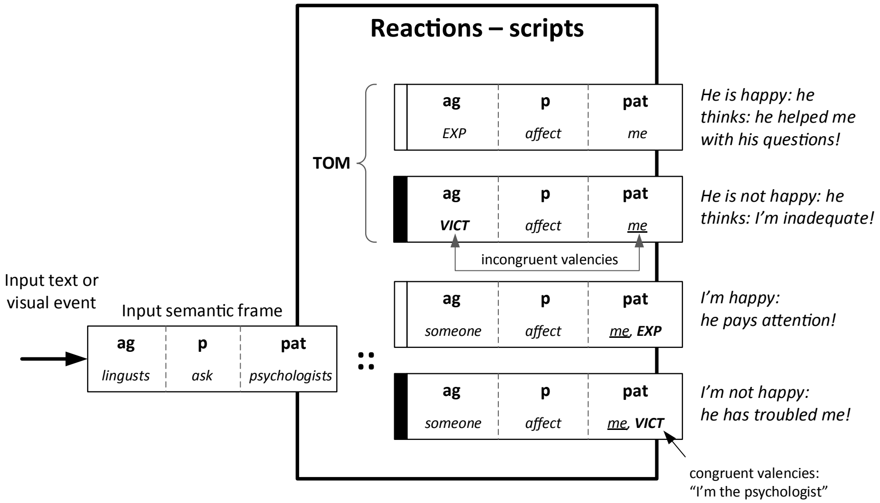

Reactions – scripts
TOM
ag p pat
EXP affect me
He is happy: he thinks: he helped me with his questions!
ag p pat
VICT affect me
He is not happy: he thinks: I'm inadequate!
incongruent valencies
Input text or visual event
Input semantic frame
ag p pat
lingusts ask psychologists
ag p pat
someone affect me, EXP
I'm happy: he pays attention!
ag p pat
someone affect me, VICT
I'm not happy: he has troubled me!
congruent valencies: "I'm the psychologist"

Fig. 4 The mechanism of the theory of mind: scripts with incongruent distribution of valencies are allocated to TOM

to be represented and considered, even if it is not the relevant representation for the input (see Fig. 5).

The other type of imagination machinery utilizes the inferences (*then*-conditions) of d-scripts. Negative scripts like DANGER provide the transfer from a negative event to the situation of aggression—'He did something bad, I have to react aggressively', or to the situation of flight—'He did something bad, I have to go away'. If the input event has been associated with the premise of such scripts (even with low relevance, as other scripts provide more relevant representation), the robot allocates the inferences of the script as a beneficial judgement in the situation: *I may consider with pleasure*

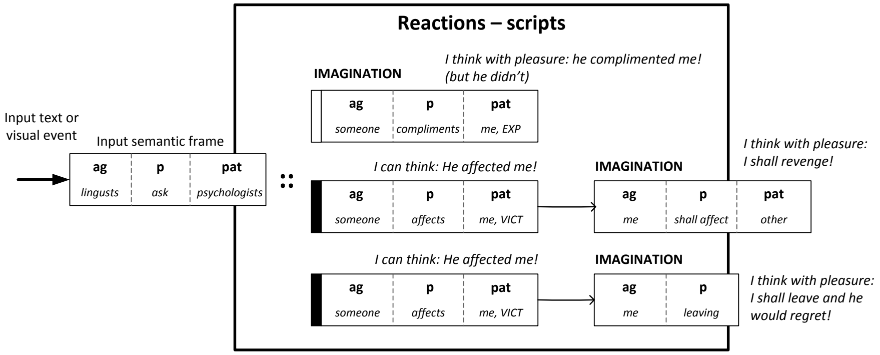

Reactions – scripts
IMAGINATION
I think with pleasure: he complimented me! (but he didn't)
ag p pat
someone compliments me, EXP
Input text or visual event
Input semantic frame
ag p pat
lingusts ask psychologists
I can think: He affected me!
ag p pat
someone affects me, VICT
IMAGINATION
ag p pat
me shall affect other
I think with pleasure: I shall revenge!
I can think: He affected me!
ag p pat
someone affects me, VICT
IMAGINATION
ag p
me leaving
I think with pleasure: I shall leave and he would regret!

Fig. 5 The mechanism of imagination: non-relevant to the situation, but emotionally relevant scripts are allocated to imagination

the possibilities to revenge (e.g. trouble linguists with some questions) or *to distance from the linguists and thus make them suffer*.

4 Conclusion

The theory of mind and imagination can be considered at extremely sophisticated cognitive mechanisms that are hard to be modelled. At the same time, they can be represented as domains, accommodating emotional judgements, that are generated during the processing of incoming stimulus, but not quite relevant to it. The relevance here is considered as the existence of another, better script with high relevance. This script suggests the view, which the organism considers as the "true" representation of events, while less relevant scripts suggest some "false" representations of the same stimulus. At the same time, these secondary, non-relevant representations can be allocated into separate mental domains, if they have a specific distribution of valencies or high emotional significance. For that, the model has to operate on a big number of representations for each stimulus, not being limited to the first representation.

Acknowledgements The research is supported by the grant of the Russian Science Foundation No. 19-18-00547, https://rscf.ru/project/19-18-00547/.

References

1. Baron-Cohen, S. (1991). Precursors to a theory of mind: Understanding attention in others. In *Natural theories of mind: Evolution, development, and simulation of everyday mindreading* (pp. 233–251). B. Blackwell.
2. Minsky, M. L. (1988). *The society of mind*. Touchstone Book.
3. Sloman, A., & Chrisley, R. (2003). Virtual machines and consciousness. *Journal of Consciousness Studies, 10*(4–5), 133–172.
4. Bocharov, D. V., Bichineva, V. S., & Granovsky, D. (2010). Open corpora: Operating principles and prospects. In *Computational linguistics and development of semantic search on the Internet: Proceedings of the scientific seminar of the XIII all-Russian United conference "Internet and Modern Society"*. Saint-Petersburg.
5. Gladky, A. V. (1985). *Syntactic structures of natural language for automation system of discourse*. Nauka.
6. Kotov, A., Zinina, A., & Filatov, A. (2015). Semantic parser for sentiment analysis and the emotional computer agents. In *Proceedings of the AINL-ISMW FRUCT 2015* (pp. 167–170).
7. Shvedova, N. Yu. (1998). *Russian semantic dictionary. RAS, Institute of Russian language*. Azbukovnik.
8. *Rusvectores*. Retrieved January 10, 2022, from https://rusvectores.org/static/models/rusvectores4/unigrams/ruwikiruscorpora-nobigrams_upos_skipgram_300_5_2018.vec.gz
9. Fillmore, J. (1968). The case for case. In E. Bach & R. T. Harms (Eds.), *Universals in linguistic theory* (pp. 1–88). Holt, Rinehart, and Winston.
10. Schank, R. C. (1975). *Conceptual information processing*. North Holland.
11. Vilhjálmsson, H., et al. (2007). The behavior markup language: recent developments and challenges. In *Intelligent virtual agents* (pp. 99–111).

12. Ochs, M., Niewiadomski, R., Pelachaud, C., & Sadek, D. (2005). Intelligent expressions of emotions. In *ACII 2005*. LNCS 3784 (pp. 707–714). Springer-Verlag.
13. Kotov, A. A. (2004). D-scripts model for speech influence and emotional dialogue simulation. In *Proceedings of the 7th Annual CLUK Research Colloquium* (pp. 134–140). University of Birmingham.
14. Baker, C. F., Fillmore, C. J., & Lowe, J. B. (1998). *The Berkeley FrameNet Project*.