



РОССИЙСКИЙ ГОСУДАРСТВЕННЫЙ ПЕДАГОГИЧЕСКИЙ УНИВЕРСИТЕТ им. А. И. ГЕРЦЕНА
HERZEN STATE PEDAGOGICAL UNIVERSITY of RUSSIA

ISSN 2687-0215

**JOURNAL
OF APPLIED LINGUISTICS
& LEXICOGRAPHY**

T. 1 № 2 2019

VOL. 1 No. 2 2019



Herzen State Pedagogical University of Russia
Российский государственный педагогический
университет им. А. И. Герцена

journall.org
ISSN 2687-0215 (online)
DOI 10.33910/2687-0215-2019-1-2
2019. Vol. 1, no. 2
2019. Том 1, № 2

Journal of Applied Linguistics and Lexicography

Mass Media Registration Certificate [EL No. FS 77 – 74246](#),
issued by Roskomnadzor on 9 November 2018
Peer-reviewed journal
Open Access
Published since 2018
2 issues per year

Свидетельство о регистрации СМИ ЭЛ № ФС 77 – 74246,
выдано Роскомнадзором 09.11.2018
Рецензируемое научное издание
Журнал открытого доступа
Учрежден в 2018 году
Выходит 2 раза в год

Editorial Board

Editor-in-chief
Sergey I. Monakhov (St Petersburg, Russia)
Deputy Editor-in-chief
David Talbot (Moscow, Russia)
Assistant Editor
Irina E. Vasileva (St Petersburg, Russia)
Renate Belentschikow (Magdeburg, Germany)
Marcello Garzaniti (Florence, Italy)
Elena V. Generalova (St Petersburg, Russia)
Tatyana Grigoryanova (Bratislava, Slovakia)
Vladimir Dubichinski (Warsaw, Poland)
Dan E. Davidson (Washington, USA)
Ekaterina V. Enikeeva (Moscow, Russia)
Maria L. Kovshova (Moscow, Russia)
Liu Limin (Beijing, China)
Olga V. Mitrenina (St Petersburg, Russia)
Sergey A. Myznikov (St Petersburg, Russia)
Marina V. Pimenova (Vladimir, Russia)
Igor V. Ruzhitskii (Moscow, Russia)
Lyudmila V. Rychkova (Grodno, Belarus)
Galina N. Sklyarevskaya (St Petersburg, Russia)
Stella N. Tseitlin (St Petersburg, Russia)
Valentina D. Chernyak (St Petersburg, Russia)
Larisa L. Shestakova (Moscow, Russia)

Publishing house of Herzen State Pedagogical
University of Russia
48 Moyka Emb., St Petersburg, Russia, 191186
E-mail: izdat@herzen.spb.ru
Phone: +7 (812) 312-17-41

Published at 02.10.2019

The contents of this journal may not be used in any
way without a reference to the journal “Journal
of Applied Linguistics and Lexicography” and the author(s)
of the material in question.

Редакционная коллегия

Главный редактор
С. И. Монахов (Санкт-Петербург, Россия)
Зам. главного редактора
Дэвид Талбот (Москва, Россия)
Ответственный секретарь
И. Э. Васильева (Санкт-Петербург, Россия)
Р. Беленчиков (Магдебург, Германия)
М. Гардзанити (Флоренция, Италия)
Е. В. Генералова (Санкт-Петербург, Россия)
Т. Григорьянова (Братислава, Словакия)
В. В. Дубичинский (Варшава, Польша)
Д. Ю. Дэвидсон (Вашингтон, США)
Е. В. Еникеева (Москва, Россия)
М. Л. Ковшова (Москва, Россия)
Лю Лиминь (КНР, Пекин)
О. В. Митренина (Санкт-Петербург, Россия)
С. А. Мызников (Санкт-Петербург, Россия)
М. В. Пименова (Владимир, Россия)
И. В. Ружицкий (Москва, Россия)
Л. В. Рычкова (Гродно, Республика Беларусь)
Г. Н. Скляревская (Санкт-Петербург, Россия)
С. Н. Цейтлин (Санкт-Петербург, Россия)
В. Д. Черняк (Санкт-Петербург, Россия)
Л. Л. Шестакова (Москва, Россия)

Издательство РГПУ им. А. И. Герцена
191186, г. Санкт-Петербург, наб. реки Мойки, д. 48
E-mail: izdat@herzen.spb.ru
Телефон: +7 (812) 312-17-41

Объем 3,03 Мб

Подписано к использованию 02.10.2019

При использовании любых фрагментов ссылка
на журнал «Journal of Applied Linguistics and Lexicography»
и на авторов материала обязательна.

Редактор *В. М. Махтина*
Редакторы английского текста *О. В. Колотина, И. А. Наговицына, А. С. Самарский*
Корректор *А. Ю. Гладкова*
Оформление обложки *О. В. Рудневой*
Верстка *А. М. Ходан*

Санкт-Петербург, 2019

© Российский государственный педагогический университет
им. А. И. Герцена, 2019

CONTENTS

APPLIED LINGUISTICS

- Monakhov S. I.* One mechanism of Russian poetic language 315
- Severskaya O. I.* Digital communication losses during
the transition from “communication” to “information transfer” 331
- Sokolov E. G.* The project of a deeply tagged parallel corpus
of Middle Russian translations from Latin 337
- Shukshina E. A.* The impact of some linguistic features
on the quality of neural machine translation 365

MODERN LEXICOGRAPHY

- Generalova E. V.* Obsolescent vocabulary of the Russian language:
Educational and lexicographic interpretation issues 371
- Sergeev M. L.* Histories, museums and dictionaries:
The forms of representing linguistic knowledge in the 16th century
and the earliest linguistic handbooks 381
- Husnutdinov A. A.* Explanatory reference dictionary as a textbook. 390

MAKING THE COMPLICATED SIMPLE

- Mitrenina O. V.* Artificial neural networks and natural language processing 399
- Sergeev M. L.* History of the written word:
On some principles of 16th century etymology 409

СОДЕРЖАНИЕ

ПРИКЛАДНАЯ ЛИНГВИСТИКА

<i>Monakhov S. I.</i> One mechanism of Russian poetic language	315
<i>Северская О. И.</i> Цифровые потери в коммуникации при переходе от «общения» к «передаче информации»	331
<i>Sokolov E. G.</i> The project of a deeply tagged parallel corpus of Middle Russian translations from Latin	337
<i>Shukshina E. A.</i> The impact of some linguistic features on the quality of neural machine translation	365

СОВРЕМЕННАЯ ЛЕКСИКОГРАФИЯ

<i>Генералова Е. В.</i> Устаревшая лексика русского языка: вопросы преподавания и лексикографической интерпретации	371
<i>Сергеев М. Л.</i> Истории, музеи, словари: формы представления знания о языках в XVI веке и первые лингвистические справочники (книги-полиглоты)	381
<i>Хуснутдинов А. А.</i> Толковый словарь-справочник как учебное пособие	390

ПРОСТО О СЛОЖНОМ

<i>Митренина О. В.</i> Нейронные сети и компьютерная обработка языка	399
<i>Сергеев М. Л.</i> История письменного слова: о некоторых принципах этимологии XVI в.	409

ONE MECHANISM OF RUSSIAN POETIC LANGUAGE

S. I. Monakhov✉¹¹ Herzen State Pedagogical University of Russia, 48 Moika River Emb., Saint Petersburg 191186, Russia

Abstract. Traditionally, the phenomenon of the semantic aura of the verse metre was regarded exclusively as historically determined; the question of a potential synaesthesia (the imitative potential possessed by the rhythmic structure of a poetic text) was essentially disregarded. This paper aims to approach the problem of “metre and meaning” from the perspective of possible actualisation of certain language forms in the metrical structures of binary and ternary metres; in other words, to analyse how the metrical nature of verse determines its basic semantic model. We have come to the conclusion that the fundamental difference between Russian binary and ternary metres lies in the level of rhythmic prominence of metrically dual words, the majority of which are pronouns. The very structure of binary metres suggests a constant possibility for pronouns to be in proximity to an unstressed syllable and to receive more or less heavy stress. In ternary metres pronouns find themselves inside the circle of metrical stresses and, being inevitably adjacent to either the preceding or the following one, lose their accent and are swallowed during pronunciation. The latter, in turn, results in weakening of deictic and anaphoric language functions and undermines the established logic of textual development. That is where different, i. e., poetic, mechanisms of creating meaning come to the fore. Ternary metres put rhythmic stress on notional words, creating — in accordance with the law of poetic analogy and via omission of intermediary elements — linguistically unpredictable associations between them; binary metres emphasise semi-notional and functional words, stressing the logical and grammatical order of text development.

Keywords: metre, meaning, Russian poetry, rhythm, pronoun.

Introduction

The problem of the interdependence of metre and meaning, which had already attracted the early reformers of the Russian verse, namely, V. Trediakovsky, M. Lomonosov, and A. Sumarokov, was formulated within the academic paradigm during the 1960s by K. F. Taranovsky (Taranovsky 2000). From 1980–2000, mainly in the works of K. D. Vishnevsky (Vishnevsky 1985), M. L. Gasparov (Gasparov 1999) and others, the phenomenon of the semantic aura of the verse metre was regarded exclusively as historically determined; the question of a potential synaesthesia (the imitative potential possessed by the rhythmic structure of a poetic text) was essentially disregarded. Only in more recent works (Khvorostianova 2014; Fridberg 2014) can we find attempts, inspired by Taranovsky’s ideas, to prove a connection between the semantics of the text and its metrical and rhythmic structure.

At the same time, the generally accepted understanding of rhythm as a result of the interaction between the prescribed metrical law of interchanging strong and weak positions, on the one hand, and the language system, on the other hand, allows us to approach the problem of “metre and meaning” from the perspective of possible actualisation of certain language forms in the metrical structures of binary and ternary metres, in other words, to analyse how the metrical nature of verse determines its basic semantic model.

The present state of Russian metrical theory categorically denies the existence of any fundamental differences in the rhythmic nature of binary and ternary metres. Cf. the characteristic fragment from an article by Gasparov:

“...nel’zya ne schitat’sya s tem, chto v soznanii chitatelej, pisatelej i teoretikov XVIII–XIX vv. (krome razve Chernyshevskogo) dvuslozhnye i trekhslozhnye razmery nikogda ne protivopostavlyalis’ kak dve sistemy: oni prinadlezhali k odnoj i toj zhe sillabo-tonicheskoj sisteme stiha, i predpolagalos’, chto v nih dejstvuyut obshchie normy ritma” [“...we cannot but take into account that in the perception of the readers, writers and scholars living in the 18th and the 19th centuries (with the only exception of Chernyshevsky perhaps), binary and ternary metres were never opposed as two separate systems: it was assumed that they belonged to the same syllabo-tonic type of verse and were governed by the same rhythmic rules”] (Gasparov 1984, 176).

Note, however, that in the early 20th century, when the actual poetic practice determined the new era theoretical constructions and was, in turn, determined by them, no one questioned the existence and significance of these differences.

Two emblematic conceptions of that period — the “logometer” (Chudovsky 1914) nature of verse and the “logaoedic” nature of prose — emerged practically simultaneously and in close connection with each other. A. M. Peshkovsky wrote about their adherents:

“Vse nazvannye avtory <...> ishchut, k sozhaleniyu, ne otlichij ritma prozy ot ritma stiha, a, naprotiv, skhodstv v etom otnoshenii prozy so stihom, t. e., v sushchnosti, annuliruyut samuyu zadachu issledovaniya, poskol’ku delo idet imenno o proze. Vse oni iskhodyat kak by iz molchalivogo predpolozheniya, chto inyh ritmicheskikh form, krome tekhn, kakie dany v stihe, byt’ ne mozhet. No s lingvisticheskoy tochki zreniya takoe predpolozhenie nichem ne mozhet byt’ opravdano” [“All these authors <...> are unfortunately looking not for the differences between the rhythm of prose and the rhythm of poetry but, quite the opposite, for similarities in this respect between them, i. e., in essence, they annul the very task of their research, since the research is about prose. All of them tacitly assume that there can be no other rhythmic forms than those that are given in verse. But this assumption, from a linguistic point of view, is completely unjustified”] (Peshkovsky 1927, 44–45).

The theoretical thought of that time, with striking persistence, rejected the traditional foot-based theory of poetic speech as an empty abstraction, totally unconnected with the real prosody¹, and at the same time kept looking for two-, three-, and four-syllable feet in prosaic speech (Bely 1919; Grossman 1928; Brodsky 1923; Shengeli 1923; Engelgardt 1923). However, this paradox is quite understandable if we take into account that the “word-rhythm” is equivalent to the “tonic” one (unit is a stress) and the “foot-rhythm” is equivalent to the “syllabic” one (unit is a group of syllables). Deconstruction of the Russian syllabo-tonic verse system at the end of the 19th century was determined by the logics of the deepening division between the domains of poetry and prose, which, in terms of prosody, were endowed with the tonic or syllabic features, respectively. Poetry started being perceived as tonic in essence.

When we say that the unit of the “word-rhythm” is stress, we do not mean just any word or any stress. We refer to the so-called “significant expressions” of verse, or — retrospectively — to the “prosodic periods”, which served as a basis for A. Kh. Vostokov’s theory of folk verse (1817). The conception of a poetic line being rhythmically arranged through semantic engineering that was born in the early modern period² reveals itself, for example, in the notorious “sdvigoloiya”

¹ Verse was regarded not as consisting of abstract rhythmic units — syllables and feet, but of real words. The war against the invariant of metrical scheme, started already by A. Bely, led in due course to L. Timofeev’s requirement “ne svodit’ bogatstvo konkretnyh stihotvornyh ritmov k sovershenno bespoleznomu i abstraktnomu metru, a nahodit’ v kazhdom sluchae tu “normu”, kotoraya sozdana dannym stilem <...> sushchestvuet li voobshche eta edinaya metrisheskaya norma, s kotoroj my sootnosim otstupayushchij ot nee ritm, kakim obrazom popal v chelovecheskoe soznanie etot ideal’nyj metrisheskij zakon?” [“not to reduce the rich variety of poetic rhythms to a completely useless and abstract metre, but find, in each case, the ‘norm’ that is created by this particular style <...> does it even exist, this universal metrical norm against which we compare the rhythm deviating from it? How did this ideal metrical law appear in the human mind?”] (Timofeev 1939, 23). See also (Chudovsky 1915; Chudovsky 1917).

² Cf. with a long forgotten, and much ridiculed in its time, idea proposed by Baron Günzburg about the *tactus*-based composition of Russian syllabo-tonic metres. Günzburg wrote in his book that “words [in ternary metres — *S.M.*] will be arranged in such a way that the stress would fall alternately on the prominent and the secondary phrase” (1915, 103).

[“shiftology”] of A. Kruchenykh. If we do not regard it as a simply ridiculous “castration of Pushkin’s work”, this academic discipline appears to be a typical phenomenon of its time, because Kruchenykh understands the shift as merging, in the process of reading, of two or more graphic words into one acoustic (phonetic) word and identifies the immanent, structural reason for the shift in the verse as the disparity between metre and language (Kruchenykh 1922; 1924). The rhythm is no longer seen as a factor of meaning distortion, but, vice versa, the meaning, or rather the sequence of “significant expressions”, is regarded as the fundamental principle of the verse’s rhythmic organisation.

The development and direction of poetic studies in the early 20th century were largely predetermined by the ongoing transformation of the poetic system itself, by the dismantling of the traditional syllabo-tonic system and the creation on its basis of a new tonic verse. It is important to remember that nobody at the time saw this process as a fundamental replacement of one structural principle with another and that during this transitional period, the impact of a gradual destruction of the old rhythmic model that was seemingly unable to accommodate the new content was extremely powerful.³ This is why the poetic perception of the period is so intriguingly “ternary”, as opposed to the “binary” perception which was characteristic of the time that saw the zenith of the Russian classical verse. Researchers proved long ago that Russian strict-stress verse emerged from ternary metres (Tomashevsky 1929; Gasparov 1968); the only regrettable thing is that too much attention was given to the analysis of the external rhythmic manifestations of this process, and not enough attention was given to the investigation of the underlying principles of the semantic construction of the verse.⁴

History of Russian ternary metres

We now turn briefly to the history of ternary metres in Russia. When ternary metres first appear, they show obvious similarity with logaoedic verse (Vishnevsky 1969a); they are still not clearly identified as the “correct” syllabo-tonic metres; and the authors of the 18th century experience specific difficulties in their application and attribution:

“I vse-taki eto bylo stanovlenie novogo i ne ochen’-to privychnogo dlya sluha togdashnego chitatelya ritmicheskogo zvuchaniya — nastol’ko neprivychnogo, chto sami avtory neredko putali trekhslozhnye razmery s logaedicheskimi formami, vo vsyakom sluchae, ohotno kombinirovali i to, i drugoe v predelakh odnogo proizvedeniya” [“And still this was also the development of a new rhythmic pattern that did not sound familiar to the readers of the time; so unfamiliar it was that authors themselves often confused ternary metres with logaoedic forms; in any case, they gladly combined both of them within one work”] (Vishnevsky 1969b, 7–8).

For example, some of G. Derzhavin’s purely dactylic poems have random amphibrachic lines, which occasionally slip in: “Ty chasto vo zerkale vodnom / Pod rdyanoy igraesh zarey, / Na zybkom lazure bezdonnom / Tenyu melkaesh tvoey...” [“Often you in the mirror of water / Play under the crimson dawn, / On the rippled, fathomless sky blueness / You flash as your own shadow”] (‘Lastochka’ [‘Swallow’]); and vice versa: “Broste svoi nedosugi, / Skachite, plyashite smelee: / Beyte v ladoshi rukami, / Shchelkayte gromko perstami” [“Cast aside your lack of free time, / Jump and dance, don’t be shy! / Clap your hands, / Snap your fingers”] (‘Lyubitelyu hudozhestv’ [‘To the lover of the arts’]).

³ Cf., for example, with the theory of a single “metrical origin” of classical and non-classical metres suggested by Bozhidar (1916) and similar ideas of S. Bobrov (1915).

⁴ For the theorists of the early 20th century, ternary metres, tonic verse and folk verse are united in a certain distinctive “semantic prosody”, as each of these systems is based on the count of “significant expressions”, usually bearing one stress. See, for example, (Bryusov 1924).

I. Bogdanovich's 'Oda duchovnaya' ['Spiritual Ode'] (V) has a secondary title 'Daktilicheskimi stihami' ['Written in dactylic verse'], even though it actually uses a regular alternation of anapaestic and amphibrachic lines.⁵

M. Kheraskov's play entitled 'Milana' contains the following line, combining lines of ternary metres with different number of feet:

“Lyubit mudrec i durak,
Lyubit sluzhitel' i barin,
Turok, Francuz i Tatarin,
Gde b ni lyubili, carstva takogo net, —
Slovom, lyubit ves' svet.
Hot' muchit lyubov',
Palit nashu krov',
No, vidno, priyatna lyubov'”.

[“Love is felt by a wise man and a fool,
Love is felt by a servant and a lord
By the Turkish man, by the French man, and by the Tatar man.
There is no such realm where people do not feel love, —
In other words, the whole world is in love.
Even though love makes us suffer,
And makes our blood burn,
Apparently, love is pleasant.”]

We can point to numerous facts to show that Russian poets of the 18th century could feel the tonic nature of ternary metres:

- (1) intermixing of lines with different anacruses and different number of feet;
- (2) sporadic intrusions of verses with different number of syllables in unstressed intervals into texts written in ternary metres; some of these inclusions are represented by lines lacking a syllable at the caesura (as in Kheraskov's example given above: “Gde b ni lyubili, tsarstva takogo net”), while some are pure tonic verses (like, for example, in Derzhavin's poem 'Na pokorenje Parizha' ['On the occasion of taking Paris']: “Vsya nas teperya vselena / Svoey uzh zashchitoy chtet; / Evropa uz svobozhdenna / Khvalnymi pesnmi poet” [“Now the whole universe / Honours us as their protection / Europe freed from chains / Sings the songs of praise”]);
- (3) predominant use of ternary metres in genres that are connected with musical performance (arias, topical songs, duets, trios, choral pieces in fantastic comedies, operas, stage plays and ballets).

The similarity between the ternary metres of the 18th century and the logaoedic verse becomes curiously manifest in the rhythmical structure of the 19th century amphibrachic trimetres. Poems of a certain meditative tone (Gasparov 1999, 120–151) that are written in this metre, as well as texts connected with the tradition of translations and imitations of Heine, show a characteristic tendency to use masculine word boundary in the first strong position of the verse —

⁵ On connection of this name with the German tradition, see (Zhirmunsky 1975).

a tendency that puts them in stark contrast with other amphibrachic works. This unusual “iambic” beginning of a line, formed by the strictly observed separation of the two first syllables, apparently emerges due to the interest of Russian authors in the German three-ictic tonic verse and its predominantly monosyllabic unstressed intervals and large number of “iambic” lines. In the work of M. A. Tarlinskaja, which analyses tonic verse in English poetry, we find the following description of the syllabic composition of a line in three-ictic tonic verse used by Heine and some other German poets of the 19th century:

“Both English and German variants of the three-ictic *dolnik* display similar features: ‘syllabo-tonic’ lines occur, on the whole, more often than in the four-ictic verse. ‘Iambic’ lines (.11) occur in about one-third to one-fifth of the lines <...> The less frequent ‘amphibrachic’ form (.22) usually takes up about another one-fifth of literary *dolnik* texts. Thus, up to one-half of the lines in the literary three-ictic *dolnik* are ‘syllabo-tonic’” (Tarlinskaja 1993, 80).

In Heine’s poems, according to Tarlinskaja, lines with two disyllabic inter-ictic intervals account for 20.4% of all verses, while the lines using the 2 + 1 scheme account for 27.2%. Taking into account that 92.7% of lines in Heine’s three-ictic tonic verse have a monosyllabic anacrusis, it seems quite possible that Russian poets, striving to recreate features typical for the tonic verse of Kotzebue, Heine and other authors, without going beyond the limits of the traditional syllabo-tonic system, preferred to place a word with a masculine ending in the first strong position of the amphibrach, where the rhythmical inertia of the line has not yet established itself, thus achieving an illusion of combining monosyllabic and disyllabic intervals within the verse — an illusion that becomes even more prominent when the first masculine word boundary is emphasized by the boundary of syntactic segments.

The scarce logaoedic verses of Russian 19th century poets were aimed at creating the same effect but with the help of completely different techniques. Thus, three of the most well-known poems written by V. Zhukovsky (‘Zhaloba pastuha’ [‘The Lament of the Shepherd’]), M. Lermontov (‘Oni lyubili drug druga tak dolgo i nezhno...’ [‘They loved each other so long and so tender...’]) and A. Fet (‘Izmuchen zhizn’yu, kovarstvom nadezhdy...’ [‘Saddened by life, the treachery of hope...’]) (Kholshevnikov 1991) follow the same pattern: (1)—1—2...[the number of disyllabic unstressed segments can differ]...2—(n). Zhirmunsky described the structure of such lines in Fet’s poem as a combination of two iambic feet and two amphibrachic feet (Zhirmunsky 1975, 213). Examination of the arrangement of word boundaries, however, suggests one iambic foot and several amphibrachic feet: the first word with a metrical stress in these logaoedic verses has a preferential masculine ending, while the second word has a feminine or dactylic one. The rhythmical effect achieved in this way is similar to the feeling that appears when we read amphibrachic verses with the first masculine word boundary. Cf., for instance, the following Zhukovsky’s lines:

- (1) logaoedic verse: “Na tu | znakomuyu goru / Sto raz | ya v den prikhozhu; / Stoyu, | sklonyasya na posokh, / Iv dol | svershiny glyazhu” [“That familiar mountain / I visit a hundred times per day; / I stand, leaning against the staff, / Looking at the valley below from the top”];
- (2) amphibrachic trimetre: “Kuda | uleteli tak skoro? / Pechal | poselilas v dushe [...] / Edva | ya uspela rastsvest, / Uzhe, | bezotradnaya, vyanu” [“Where did you fly so fast? / Sadness has found its home in my heart [...] / Barely had I the time to bloom, / Already, miserable, I am withering”].

In the early 20th century, and primarily in B. Pasternak’s work, “tonic” lines in which monosyllabic and disyllabic unstressed intervals can be found side by side start randomly sneaking into

poems written in ternary metres (and especially in amphibrachic trimetres).⁶ We find 23 such cases in Pasternak's amphibrachs, with 22 of them representing the pattern (1)—2—1. Only in three cases the first word boundary does not come after the fifth syllable. In other examples, we see either a masculine ending of the second metrically stressed word—in cases when the verse does not have tribrachs: “Nad shabashem skal, | k kotorym...” [“Over the orgy of rocks, in which...”], “I vdrug — iz sadov, | gde tvoy...” [“And suddenly — from the gardens where your...”], “Lish glaz nocheval, | iz milogo...” [“Only your eye spent the night, from the sweet...”], or a hyperdactylic ending of the first metrically stressed word—in cases when there is a tribrach in the second foot: “I vypryamitsya, | kak prezhdde...” [“And straighten up as before...”], “Tseluyushchikhsya | i pyushchikh...” [“Kissing and drinking...”], “Okliknutye | s pozitsiy...” [“Hailed from the positions...”].

In practically all fully stressed lines, as we can see in the examples above, there is a syntactic boundary after the second strong position. These three factors — the masculine word boundary, the syntactic division, and the omission of the unstressed syllable — act together to break the verse into two uneven parts. The second masculine word boundary in Pasternak's amphibrach sounds like a strong rhythmical break disrupting the prescribed inertia of the verse; the disruption is so strong that it also justifies, as it were, the sporadic shortening of the unstressed interval following it. (Cf. this technique with the verse structure of the “Heine-type” amphibrachic trimetre using a distinct “iambic” beginning, which sounds similar to some of the 19th century logaoedic verses.) This is easily noticeable if we compare pairs of adjacent lines using a masculine word boundary in the first and the second strong position of the verse, with one of the lines being “tonic” in type:

- (1) “Spit stroy | sosnovykh vysot, / I les | shelushitsya i kaplyami [...]” [“Sleeping are the lines of high pines, / And the forest is exfoliating, and the drops [...]”];
- (2) “Glukhaya pora | listopada, / Poslednikh gusey | kosyaki. / Rasstraivatsya | ne nado: / U strakha glaza | veliki” [“The dead season of the leaf drop / The last flocks of geese. / There is no need to be upset: / Fear sees danger everywhere”].

Unlike with the Russian poetry of the first half of the 19th century, we should not be surprised to find the rhythmical “break” of the amphibrachic trimetre in Pasternak's poetry predominantly on the second strong position of the verse. If in the early 19th century it was important to prevent the emergence of the rhythmic inertia, in the late 19th and early 20th centuries, it was important to disrupt it.

Preferences in the organisation of the several “tonic” inclusions that we find in Pasternak's amphibrachic trimetre and — much less frequently — his anapaestic trimetre are fully in harmony with the basic rhythmical tendencies that Gasparov identified in the Russian three-ictic dolnik of the 20th century:

- (1) first, the gradual strengthening and, already by the 1930s–1940s, the dominance of form III (anacrusis 2/0, interval 2—1; “Ia pomnyu stupeni trona” [“I remember the steps of the throne”]) and form V (anacrusis 2/0, interval 4; “Neozhidanniy akvilon” [“Unexpected aquilon”]);
- (2) second, the preferential use of the masculine word boundary in the last monosyllabic interval (Gasparov 1968, 71–72, 83–84).

⁶ Nevertheless, these poems remain distinctly anapaestic or amphibrachic: the incursions of “tonic” lines are apparently accidental in character. See Gasparov, “...it is difficult, for instance, to describe as tonic a large poem that is written from beginning to end in a regular anapaest and only in one or two verses permits a monosyllabic interval instead of a disyllabic one: such poem will be perceived as a regular anapaest with a random deviation” (Gasparov 1968, 71).

The rhythmic preferences of ternary trimetres⁷ in Pasternak's and, to a lesser extent, Blok's work, as well as possibly in the works of other poets of the 20th century⁸, played an important role in the development of the Russian three-ictic *dolnik*. The mechanism of rhythm formation in both systems was the same: creation of the rhythmic inertia of the line and its subsequent destruction. The difference was in the means used for bringing it into action: either by changing the feminine (dactylic) word boundary to a masculine one in ternary metres or by "shortening" the second unstressed interval in tonic verse. The contrasting rhythmic structure of the amphibrachic line, based on the dissimilarity of word boundaries, in the early modern period was no longer perceived as aesthetically significant and expressive. The updating, "enlivening" of this model was achieved by means of omitting one unstressed syllable in a corresponding inter-stress interval.

Various instances of metrical stress omissions, "dropping" unstressed syllables, and blending lines with anacruses of unequal numbers of syllables clearly indicate that Pasternak's ternary metres start revealing the "tonic" constituent of their rhythm. In other words, perception of the tonic essence of ternary metres, which was characteristic of the Russian 18th century poets, had been, in a sense, revived by the 1910s–1930s. However, while logaoedic and tonic verses of Derzhavin and his followers get "evened out" to form the regular ternary metres, by the 20th century, tonic verses began to supplant the ternary metres that gave birth to them. The principles of "semantic versification", developed within the traditional, rigidly regulated system possessing only a limited inventory of rhythmic resources, are now awake and imperiously breaking up the cradle that has become too small for them. Although another matter altogether, it is worth noting that only within such a strictly regulated system could these principles have emerged and gained strength.

Pronouns as words of dual metrical natures

A significant step concerning the problem of correlation between prosody and metre was made by Russian poetic science in the 20th century, when a clear distinction between stressed and unstressed words was made and a special type of words with dual metrical natures were recognized.⁹ Zhirmunsky, who should be credited with these advances, counted among these dual metrical words such word classes as pronouns, pronominal adverbs and conjunctions, monosyllabic numerals, auxiliary verbs and interjections, and pointed out the specific dual character of their accentuation: all of these words become unstressed in immediate proximity to a stressed syllable, but retain a more or less noticeable accent when they are adjacent to an unstressed syllable.

Unfortunately, the sophisticated method for differentiating the weight of stress falling on metrically dual words, which was developed by Zhirmunsky, as well as his principle of regarding stress as a system of certain quantitative relations, never took hold in poetic theory — both because of its complexity, which inevitably led to the excessive fragmentation of statistical indicators, and

⁷ The strongest influence here is perhaps exercised by the amphibrach, since its metrical structure makes it inherently prone to be replaced by tonic verse. Russian accentual verse emerged slowly and tentatively, justifying itself at the early stage of development simply by the uneven number of syllables in unstressed intervals; and it can be seen as lucky when such a possibility was suggested by the scheme of a classical metre.

⁸ Cf., for example, with similar rhythmical tendencies that O. A. Orlova finds in A. Tvardovsky's amphibrach (Orlova 1985).

⁹ Cf. with the following: "Spornymi dlya russkoj prosodii yavlyayutsya akcentnye otnosheniya osoboj kategorii maloudarnyh slov, po preimushchestvu — odnoslozhnyh, rezhe — dvuslozhnyh, kotorye zanimayut kak by srednee polozhenie mezhdru slovami znachashchimi (ponyatiyami) i slovami chisto sluzhebnyimi (kak predlogi i soyuzy): syuda otnosyatsya, glavnyim obrazom, mestoimeniya i nekotorye narechiya (mestoimennyye), vspomogatel'nye glagoly i nemnogie drugie" ["A controversial aspect of Russian prosody is accentual relations of a special group of lightly-stressed words, predominantly monosyllabic and more rarely disyllabic, which stand in the middle position between the notional words (concepts) and purely functional words (such as prepositions and conjunctions): they include, first and foremost, pronouns and some adverbs (pronominal adverbs), auxiliary words and a few others" (Zhirmunsky 1975, 87).

due to the lack of linguistic experimental data that would prove the hypothesis suggested by Zhirmunsky. Such data are still absent today, and we do not know if it will ever be available or even if it is at all necessary. For the practical purposes of scansion, scholars have been successfully using a set of standard operations that were first outlined by Zhirmunsky and later fully developed by Gasparov and T. V. Skulacheva (Gasparov, Skulacheva 2004).

For poetic theory, the fruitfulness of Zhirmunsky's innovation lies in the understanding of the conventional nature of poetic stress. Special laws of accentuation govern the verse; the sequence of words is laid over a predetermined rhythmic frame with mapped out strong and weak positions, so that the accentual movement is adjusted to the prescribed pattern. In this sense, it would be fruitless to compare the verse with everyday speech and general language laws. Poetic speech distorts the accent of the word as much as it distorts its sound composition.¹⁰ It is important to identify the laws governing this deformation, and then, after comparing the rhythmic structure of binary metres loosened by the frequent omissions of stress, on the one hand, and the rigid rhythmic pattern of ternary metres created by fixed tonic constants, on the other, we will get a clearer understanding of the fundamental differences between these metres — differences that ultimately determined their different fates in the history of Russian poetry. The crucial point here is the metrically dual words, the majority of which are pronouns.

The level of rhythmical prominence of pronouns in binary and ternary metres is necessarily different. Frequent dibrachs in iamb and trochee create many more possibilities for stressing pronouns; the very structure of binary metres suggests a constant possibility for pronouns to be in proximity to an unstressed syllable and, according to Zhirmunsky's theory, to receive more or less heavy stress. The situation is different in ternary metres, where pronouns find themselves inside the circle of metrical stresses and, being inevitably adjacent to either the preceding or the following one, lose their accent and are swallowed during pronunciation. The only exception here — the first syllable of the anapaest — does not play such a significant role because it stands in anacrusis, i. e., is extrametrical.

It is telling to compare, for instance, N. Nekrasov's poem 'Trojka' ['Trio of horses'], where the frequent use of "ty/tvoy" ["you/your"] in unstressed positions creates an impression of a special technique of depersonalisation, removal of the addressee:

- (1) "Chto **ty** zhadno glyadish na dorogu" ["Why are you looking so hard at the road"], "Vsyo litso **tvoye** vspykhnulo vdruzg" ["The whole face of yours suddenly flashed"], "I zachem **ty** bezhish toroplivo" ["And why are you running in a hurry"], "Polyubit **tebya** vsyakiy ne proch" ["Anyone would be happy to love you"], "V volosakh **tvoikh**, chernykh kak noch" ["In your hair, black like the night"], "Skvoz rumyanes shcheki **tvoey** smugloy" ["Through the bloom of your swarthy cheek"], "Da ne to **tebe** palo na doly" ["But you were to get a different lot"], "Budet bit **tebya** muzhpriverednik" ["You will be beaten up by your demanding husband"], "Pogruzishsya **ty** v son neprobudnyi" ["You will plunge in a sleep without waking"], "I v litse **tvoyem**, polnom dvizhenya" ["And in your fact that is full of motion"], "Kak proydesht **ty** tyazhelyi svoy put" [As you will walk your difficult road], "Ne nagnat **tebe** beshenoy troiki" ["You will not catch up with the crazy-paced horses"], —

with his love poems, where the same pronouns are placed in the metrically strong positions:

- (2) "Ya ne lyublyu ironii tvoey" ["I do not like the irony of yours"], "A nam s toboy, tak goryacho lyubivshim" ["And it's for me and you, who loved so passionately"],

¹⁰ Cf. with Jakobson's commentary on the "dualism of the prescribed and actual series of stresses" in syllabo-tonic verse, rhythmical inertia of which prescribes artificial phrasing and intonation to poetic speech and changes the rhythm of the word by giving prominence to unstressed syllables (Jakobson 1923, 101–112).

“Svidanie prodlit zhelaesh ty” [“You want to continue the meeting”], “Kak ty krotka, kak ty poslushna” [“You are so gentle, so docile”], “Ya posetil tvoe kladbishche” [“I have visited your grave”], “I obraz tvoy svetley i chishche” [“And your image is more fair and pure”], “Vstrechalsya grustno ya s toboy” [“I used to meet you with sadness”], “Ni smekh, ni govor tvoy veselyy” [“Neither your laughter nor your merry speech”], “Zabudus, ty peredo mnoyu” [“I fall into reverie, and here you are in front of me”].

This hypothesis lends itself to statistical testing. If the role of pronouns in Russian binary and ternary metres is as different as we assume, it should result in their contrasting distribution across weak and strong metrical positions. One would expect to find an overrepresentation of pronouns in the strong positions in binary metres and in the weak positions in ternary metres. To test this assumption, we calculated the raw frequencies of all monosyllabic personal pronoun forms¹¹ in the poetic subcorpus of the Russian National Corpus, both for binary and ternary metres (7,266,779 and 1,202,535 words, respectively). The results are provided in Table 1.

Table 1. Distribution of personal pronouns in Russian binary and ternary metres

Word forms	Ternary metres					Binary metres				
	Overall	Weak		Strong		Overall	Weak		Strong	
	Raw freq.	Raw freq.	%	Raw freq.	%	Raw freq.	Raw freq.	%	Raw freq.	%
ya ‘I’	16455	14231	25.49	2224	3.98	93480	54218	15.55	39262	11.26
ty ‘you’	8189	6480	11.61	1709	3.06	50102	26222	7.52	23880	6.85
on ‘he’	6640	5745	10.29	895	1.60	46943	27823	7.98	19120	5.48
my ‘we’	4365	3668	6.57	697	1.25	21868	11619	3.33	10249	2.94
vy ‘you’	1555	1203	2.15	352	0.63	10257	5122	1.47	5135	1.47
mne ‘I-DAT’	6268	4698	8.41	1570	2.81	35839	16670	4.78	19169	5.50
mnoj ‘I-INSTR’	739	29	0.05	710	1.27	4815	170	0.05	4645	1.33
im ‘he-INSTR / they-DAT’	922	727	1.30	195	0.35	6231	2999	0.86	3232	0.93
nem ‘he-PREP’	682	295	0.53	387	0.69	5142	1564	0.45	3578	1.03
ej ‘she-DAT’	949	754	1.35	195	0.35	6438	3000	0.86	3438	0.99
nej ‘she-PREP’	1012	363	0.65	649	1.16	7539	1603	0.46	5936	1.70
nas ‘we-PREP’	1567	751	1.35	816	1.46	10964	2705	0.78	8259	2.37
nam ‘we-DAT’	1755	1359	2.43	396	0.71	12051	5323	1.53	6728	1.93

¹¹ It does not seem reasonable to take into account forms with more than 1 syllable, since they will inevitably stretch over both weak and strong metrical positions in binary metres.

Table 1. Distribution of personal pronouns in Russian binary and ternary metres (continued)

vas 'you-ACC'	697	328	0.59	369	0.66	5893	1435	0.41	4458	1.28
vam 'you-DAT'	701	476	0.85	225	0.40	5675	2499	0.72	3176	0.91
ih 'they-ACC'	2608	2201	3.94	407	0.73	19788	9793	2.81	9995	2.87
nih 'they-PREP'	725	247	0.44	478	0.86	5656	955	0.27	4701	1.35
TOTAL	55829	43555	78.02	12274	21.98	348681	173720	49.82	174961	50.18

We have found a significant association between the metre and the metrical position of personal pronouns: $\chi^2(1) = 15384$, $p < 0.001$. The odds of a personal pronoun being used in a weak position in ternary metres are 3.5 times greater than those in binary metres (Fig. 1). Thus, our hypothesis is confirmed.

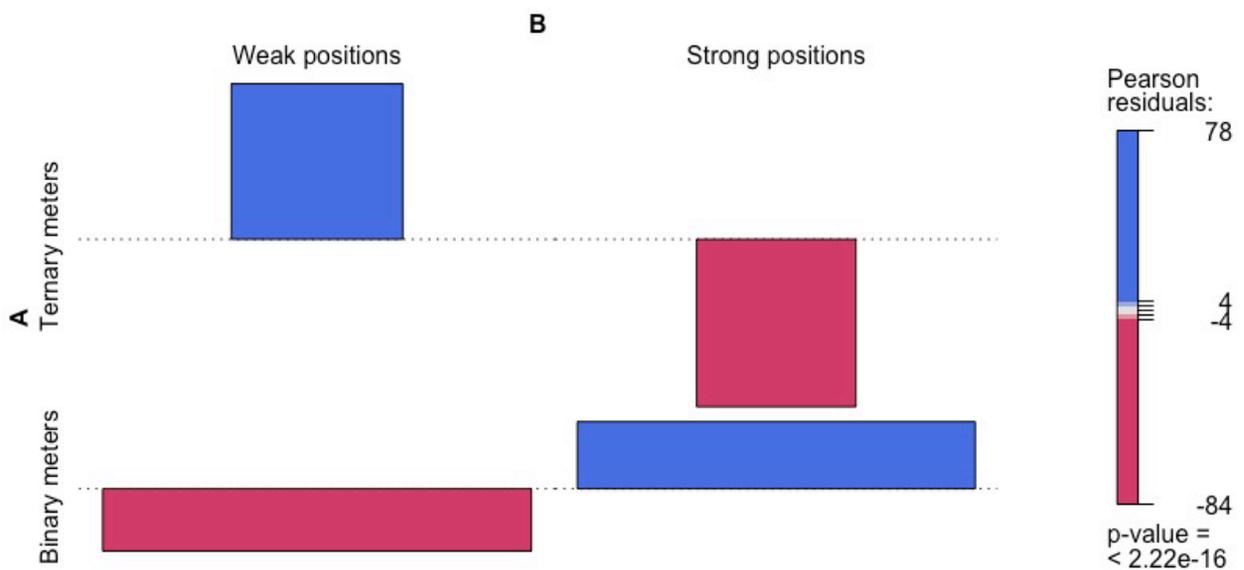


Fig. 1. Association plot of personal pronouns on weak and strong positions in Russian binary and ternary metres

However, having fitted a linear regression model to the data, with log-transformed frequency of personal pronouns as dependent variable and with metre and position as independent variables, we got the following results (Table 2):

Table 2. Coefficients of the Model.1

	B (SE)	95% CI	
		Lower	Upper
Constant	8.92***(0.28)	8.35	9.49
Metre	-2.59*** (0.40)	-3.4	-1.79
Position	-0.54 (0.40)	-1.34	0.26
Interaction term	1.15* (0.57)	0.01	2.29

Significance codes: *** — $p < 0.001$, * — $p < 0.05$.

Overall, Model.1 was highly significant: LR $\chi^2(3) = 41.7$, $p < 0.001$ and explained substantial amount of variance: pseudo- $R^2 = 0.45$ (Cox and Snell), 0.46 (Nagelkerke). The null hypothesis that the deviance of the model does not differ from the deviance of a model without any predictors can be rejected. However, of the two predictors, only one, namely metre, has crossed the threshold of statistical significance ($p < 0.001$). As for the metrical position, it has failed to make any significant contribution to the regression model ($p = 0.89$), which is confirmed by the fact that its confidence interval based on the bootstrap crosses zero. Nevertheless, the interaction term of metre and position was significant, suggesting that some other factors should be taken into consideration.

With this in mind, we fitted another linear regression model to the data, this time with the following independent variables: 1) metre (binary or ternary), 2) position (weak or strong), 3) case (nominative or non-nominative), 4) person (first, second, and third), and 5) number (singular or plural) and all possible interactions between them.

Model.2 was also highly significant: LR $\chi^2(15) = 87.75$, $p < 0.001$ and explained a much greater amount of variance as compared to Model.1: pseudo- $R^2 = 0.72$ (Cox and Snell), 0.74 (Nagelkerke).

We used the *drop1()* function in RStudio (The R Project for Statistical Computing 2013) to remove each term from the Model.2, one at a time, and test the changes in the model's fit. The results showed that only interactions of 1) metre and position and 2) position and case make significant contribution to the model, while all other variables and interactions can be easily left out. That was confirmed by the stepwise backwards model selection based on Akaike's information criterion (AIC) with the help of the *step()* function in RStudio.

Having left only these two interactions, we fitted the Model.3 to the data. Comparing Model.2 and Model.3 with the help of *anova()* function in RStudio proved that Model.3, despite the greatly reduced number of predictors, was not any worse than Model.2 ($p = 0.7$). The summary of the Model.3 is given in the Table 3:

Table 3. Coefficients of the Model.3

	B (SE)	95% CI	
		Lower	Upper
Constant	9.56*** (0.32)	8.91	10.21
Metre:ternary	-2.59*** (0.31)	-3.22	-1.97
Position:weak	0.3 (0.46)	-0.61	1.23
Case:non-nominative	-0.91** (0.34)	-1.59	-0.22
Metre:ternary : Position:weak	1.15* (0.44)	0.27	2.03
Position:weak : Case:non-nominative	-1.2* (0.48)	-2.16	-0.23

Significance codes: ***— $p < 0.001$, **— $p < 0.01$, *— $p < 0.05$.

Summary: LR $\chi^2(5) = 78.95$, $p < 0.001$; pseudo- $R^2 = 0.68$ (Cox and Snell), 0.7 (Nagelkerke).

These effects clearly show that distribution of personal pronouns in binary and ternary metres differs not only with regard to strong and weak metrical positions but also with regard to the pronouns' preferred case forms. To support this observation, we performed two independent one-tailed t-tests with Welch's correction to compare 1) the mean log-transformed frequencies of nominative personal pronouns on weak and strong metrical positions in ternary metres and 2) the mean log-transformed frequencies of non-nominative personal pronouns on weak and strong metrical positions in binary metres.

The first test revealed that, on average, the nominative pronouns on weak metrical positions in ternary metres produced significantly greater values ($M = 8.45$, $SE = 0.33$) than the nominative pronouns on strong metrical positions in the same metres ($M = 6.87$, $SE = 0.15$), $t(7) = 3.04$, $p < 0.01$. The effect size was large: $r = 0.74$ (Field, Miles, Field 2012, 58).

Conversely, the second test proved that, on average, the non-nominative pronouns on weak metrical positions in binary metres produced significantly lower values ($M = 7.76$, $SE = 0.33$) than the non-nominative pronouns on strong metrical positions in the same metres ($M = 8.61$, $SE = 0.15$), $t(15) = -2.27$, $p = 0.01$. The effect size was also substantial: $r = 0.5$. Respective bar plots with 95% confidence intervals are provided in the Figures 2 and 3.

As for non-nominative forms of personal pronouns in ternary metres and nominative forms of personal pronouns in binary metres, they reveal no predisposition to either weak or strong metrical positions, as confirmed by another pair of t-tests with Welch's correction ($t(15) = 0.51$, $p = 0.3$ for ternary metres; $t(7) = 0.33$, $p = 0.62$ for binary metres).

Conclusion

What does a text lose when it “loses” pronouns?¹² What language functions in the text are weakened? Apparently, two of these functions are: the deictic function (reference to the participants of a particular speech act: the speaker, the listener, as well as the object pointed at by the speaker; reference to the speech situation, expression of the presupposition concerning the existence of the object in the perception of the speaker and the listener, correlation with a particular referent) and the anaphoric function (reference to previous or subsequent positions of the text) (Krylov, Paducheva 1990; Seliverstova 1988; Paducheva 1985; Otkupshchikova 1984; Levin 1973). The text that ignores pronouns deliberately resists its correlation with a specific context of speech

¹² Removal of stress from pronouns together with giving prominence to other, notional, words inevitably leads to “swallowing” of the former during pronunciation and, hence, to faster reading.

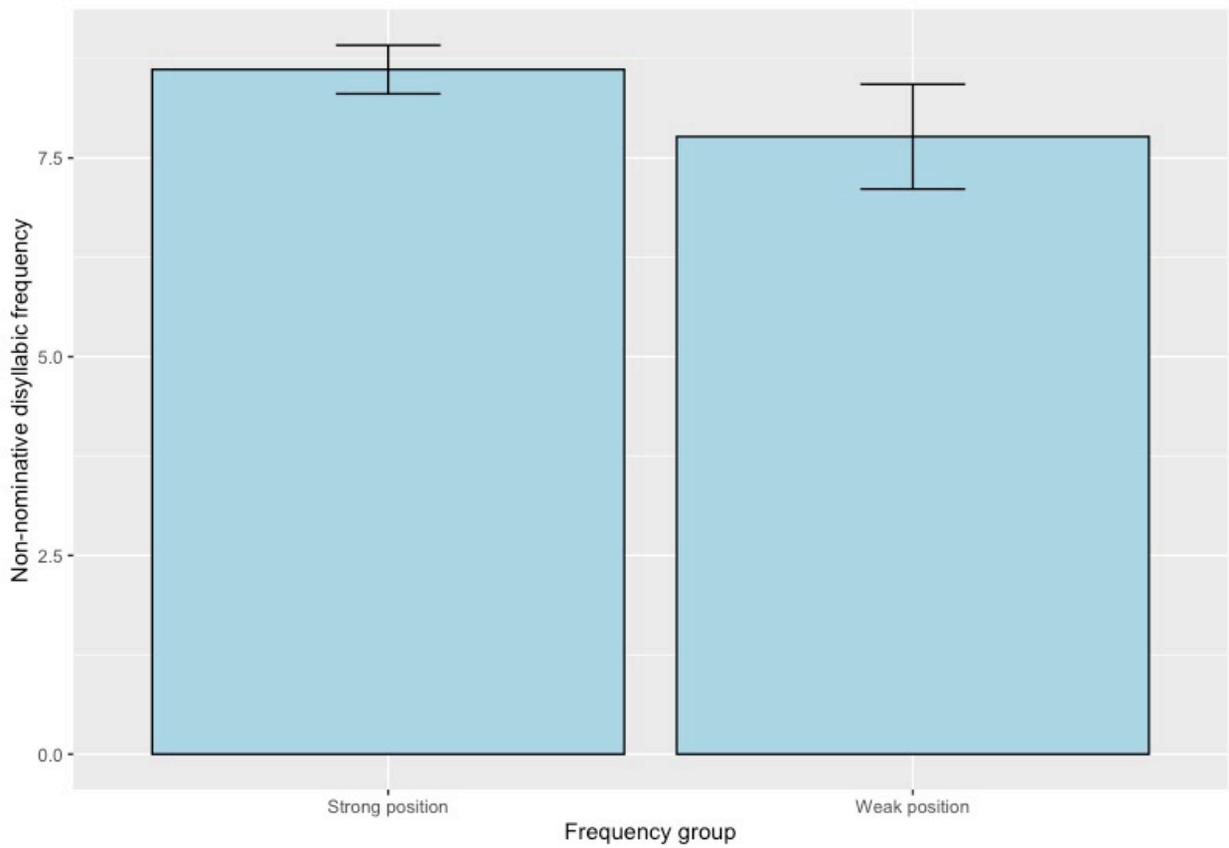


Fig. 2. Bar plots of the log-transformed frequencies of non-nominative personal pronouns on weak and strong positions in Russian binary metres

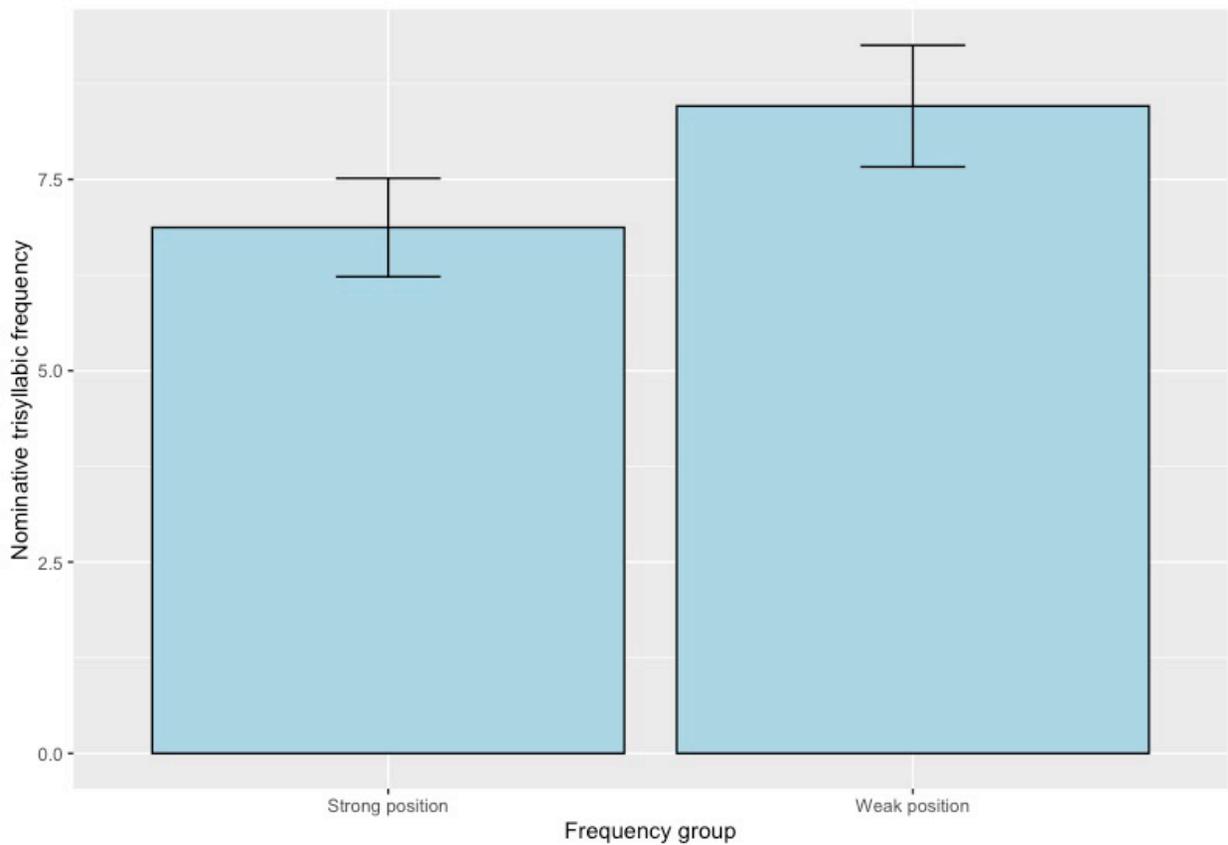


Fig. 3. Bar plots of the log-transformed frequencies of nominative personal pronouns on weak and strong positions in Russian ternary metres

situation, with extra-linguistic reality¹³, and destroys its own coherence, the consistence in the logic of representing thoughts. The text that gives prominence to pronouns and puts them in strong metrical positions does the opposite.

Both the loss of connection with reality ('defamiliarisation') and the disruption of familiar word combinations, as Russian formalists showed in their time, are among the most important features of a work of art and form the basis for the development of poetic language. Downplaying the weight of pronouns certainly does not turn a poem into a random series of sounds; it weakens the established logic of textual development and changes the relations between the words and between the text and reality. Thus, different, i. e., poetic, mechanisms of creating meaning come to the fore. Ternary metres put rhythmic stress on notional words, creating — in accordance with the law of poetic analogy and via omission of intermediary elements — linguistically unpredictable associations between them¹⁴; binary metres emphasise semi-notional and functional words (even those that are formally "unstressed", simply because of their frequent use in strong metrical positions of the verse), underlining the logical and grammatical order of text development.

Thus, ternary metres prove to be in a sense more "poetic" than binary ones; their very structure provides optimal opportunities for creating associative links (crystallisation of the "aesthetic meaning" of the word) and contributes more to the artistic effect. That is why Russian ternary metres and tonic verse are so similar to folk songs, lyrical songs, etc.: each of these accentual systems is essentially based on the count of "significant expressions" (Gasparov 1975, 77–107). In this sense, the evolution of the Russian literary verse represents a movement towards the folk verse, as it was predetermined in the 18th century.

References

- Bely, A. (1919) O khudozhestvennoy proze. In: *Gorn*. Books II–III. Moscow: Moskovskij Proletkul't Publ., pp. 49–55. (In Russian)
- Bobrov, S. P. (1915) *Novoe o stikhoslozhenii A. S. Pushkina*. Moscow: Musaget Publ., 39 p. (In Russian)
- Bozhidar [Gordeev, B. P.]. (1916) *Raspevochnoe edinstvo vsekh razmerov*. Moscow: Tsentrifuga Publ., 84 p. (In Russian)
- Brodsky, N. L. (1923) Proza "Zapisok Okhotnika". In: *Turgenev i ego vremya*. Coll. 1. Moscow; Petrograd: Gosudarstvennoe izdatel'stvo, pp. 193–199. (In Russian)
- Bryusov, V. (1924) *Osnovy stikhovedeniya [Foundations of poetic theory]*. Moscow. (In Russian)
- Chudovsky, V. (1914) O ritme pushkinskoj "Rusalki" (Otryvok). *Apollon*, 1–2: 108–121. (In Russian)
- Chudovsky, V. (1915) Neskol'ko myslej k vozmozhnomu ucheniyu o stikhe (s primernym razborom stikhoslozheniya v I glave *Evgeniya Onegina*). *Apollon*, 8–9: 55–95. (In Russian)
- Chudovsky, V. A. (1917) Neskol'ko utverzhdenij o russkom stikhe. *Apollon*, 4–5: 58–69. (In Russian)
- Engelhardt, N. A. (1923) Melodika turgenevskoj prozy. In: N. L. Brodskiy (ed.). *Tvorcheskij put' Turgeneva*. Petrograd: Seyatel' Publ., pp. 9–63. (In Russian)
- Field, A., Miles, J., Field, Z. (2012) *Discovering statistics using R*. London; Thousand Oaks, CA: Sage, XXXIV, 957 p. (In English)
- Fridberg, N. (2014) Ritm, smysl i situativnyi oreol: Belyi, Khodasevich, Tsvetaeva, Brodskii [Rhythm, sense and situational aura: Bely, Khodasevich, Tsvetaeva, and Brodsky]. In: B. P. Scherr, J. Bailey, V. T. Johnson

¹³ Somebody might disagree: regardless of the pronouns, the connection with extra-linguistic reality is preserved by the notional words of a poetic text. This is not entirely so. Language reflects reality in the entirety of all its elements. Withdrawing the elements that ensure the internal coherence of the system leads to its instability and unpredictability. It becomes easier to create new connections, which is achieved via incorporation of language elements into a special poetic structure.

¹⁴ Even the few examples of Pasternak's lines given above, when taken out of context, surprise us by their lack of coherence and create the impression of a random choice of words. The individuality of poetic style clearly reveals the systemic destruction of the mechanisms used to ensure correlation between different parts of poetic structure. Cf. with Jakobson's concept of the "differentiating rhythmic inertia" of a new verse, which significantly increases the separate value of each word stress and leads to the emergence of a special syntax; following its laws, the word group, which in ordinary speech is governed by one heavy stress, gets broken into separate, equally independent words (Jakobson 1923, 101–112).

- (eds.). *Poetry and poetics: A centennial tribute to Kirill Taranovsky. Papers of a Conference held at Dartmouth College in honor of Kirill Taranovsky*. Bloomington, IN: Slavica Publishers, pp. 215–236. (In Russian)
- Gasparov, M. L. (1968) Russkij trekhudarnyj dol'nik XX veka. In: V. M. Zhirmunsky, D. S. Likhachev, V. E. Kholshevnikov (eds.). *Teoriya stikha*. Leningrad: Nauka Publ., pp. 59–106. (In Russian)
- Gasparov, M. L. (1975) Russkij narodnyj stikh v literaturnykh imitatsiyakh. *International Journal of Slavic Linguistics and Poetics*, 19: 77–107. (In Russian)
- Gasparov, M. L. (1984) Eshche raz k sporam o russkoj sillabo-tonike. In: D. S. Likhachev, A. M. Panchenko, V. E. Kholshevnikov (eds.). *Problemy teorii stikha*. Leningrad: Nauka Publ., pp. 174–178. (In Russian)
- Gasparov, M. L. (1999) *Metr i smysl: Ob odnom iz mekhanizmov kul'turnoj pamyati*. Moscow: Russian State University for the Humanities Publ., 297 p. (In Russian)
- Gasparov, M. L., Skulacheva, T. V. (2004) *Stat'i o lingvistike stikha*. Moscow: Yazyki slavyanskoj kul'tury Publ., 283 p. (In Russian)
- Grossman, L. P. (1928) *Sobranie sochinenij v 5 t. T. 3: Turgenev: Etyudy o Turgeneve. Teatr Turgeneva*. Moscow: Sovremennye problemy Publ., 256 p. (In Russian)
- Günzburg, D. G. (1915) *O russkom stikhoslozhenii: Opyt izsledovaniya ritmicheskogo stroya stikhotvorenij Lermontova*. Petrograd: Tipografiya tovarishchestva M. O. Vol'f Publ., LXIV, 269 p. (In Russian)
- Jakobson, R. (1923) *O cheshskom stikhe preimushchestvenno v sopostavlenii s russkim*. Berlin: Gosudarstvennoe izdatel'stvo RSFSR, 120 p. (In Russian)
- Kholshevnikov, V. E. (1991) *Stikhovedenie i poeziya*. Leningrad: Leningrad State University Publ., 254 p. (In Russian)
- Khvorostianova, E. V. (2014) Metr i smysl: Postanovka problemy i istoriya ee izucheniya. In: B. P. Scherr, J. Bailey, V. T. Johnson (eds.). *Poetry and poetics: A centennial tribute to Kirill Taranovsky. Papers of a conference held at Dartmouth College in honor of Kirill Taranovsky*. Bloomington, IN: Slavica Publishers, pp. 249–260. (In Russian)
- Kruchenykh, A. (1922) *Sdvigologiya russkogo stikha: Trakhtat obizhal'nyj (Traktat obizhal'nyj i pouchal'nyj)*. Moscow: Tipografiya TsIT Publ., 46 p. (In Russian)
- Kruchenykh, A. (1924) *500 novykh ostrot i kalamburov Pushkina*. Moscow: Tipografiya TsIT Publ., 71 p. (In Russian)
- Krylov, S. A., Paducheva, E. V. (1990) Mestoimenie. In: V. N. Yartseva (ed.). *Lingvisticheskij entsiklopedicheskij slovar'*. Moscow: Sovetskaya entsiklopediya Publ., pp. 294–295. (In Russian)
- Levin, Yu. I. (1973) O semantike mestoimenij. In: A. A. Zaliznyak (ed.). *Problemy grammaticeskogo modelirovaniya*. Moscow: Nauka Publ., pp. 108–121. (In Russian)
- Orlova, O. A. (1985) Svoeobrazie ritmicheskogo dvizheniya v trekhslzhnykh razmerakh A. T. Tvardovskogo (na primere trekhstopnogo amfibrakhiya). In: L. I. Timofeev (ed.). *Russkoe stikhoslozhenie: Traditsii i problemy razvitiya*. Moscow: Nauka Publ., pp. 301–306. (In Russian)
- Otkupshchikova, M. I. (1984) *Mestoimeniya sovremennogo russkogo yazyka v strukturno-semanticheskom aspekte*. Leningrad: Leningrad State University Publ., 87 p. (In Russian)
- Paducheva, E. V. (1985) *Vyskazyvanie i ego sootnesennost' s dejstvitel'nost'yu: Referentsial'nye aspekty semantiki mestoimenij*. Moscow: Nauka Publ., 271 p. (In Russian)
- Peshkovsky, A. M. (1927) Printsipy i priemy stilisticheskogo analiza i otsenki khudozhestvennoj prozy. In: M. A. Petrovsky (ed.). *Ars poetica*. Vol. 1. Moscow: GAKhN Publ., pp. 29–68. (In Russian)
- Seliverstova, O. N. (1988) *Mestoimeniya v yazyke i rechi*. Moscow: Nauka Publ., 151 p. (In Russian)
- Shengeli, G. A. (1923) O ritmike turgenevskoj prozy. In: G. A. Shengeli. *Traktat o russkom stikhe*. Moscow; Petrograd: Gosudarstvennoe izdatel'stvo Publ., pp. 178–181. (In Russian)
- Taranovsky, K. (2000) O vzaimootnoshenii stikhotvornogo ritma i tematiki. In: K. Taranovsky. *O poezii i poetike*. Moscow: Yazyki russkoj kul'tury Publ., pp. 372–403. (In Russian)
- Tarlinskaja, M. (1993) *Strict stress-meter in English poetry compared with German and Russian*. Calgary, Alberta: University of Calgary Press, XXII, 261 p. (In English)
- The R Project for Statistical Computing*. [Online]. Available at: <http://www.R-project.org/> (accessed 01.08.2019). (In English)
- Timofeev, L. I. (1939) *Teoriya stikha*. Moscow: Goslitizdat Publ., 232 p. (In Russian)
- Tomashevsky, B. V. (1929) *O stikhe*. Leningrad: Priboj Publ., 324 p. (In Russian)
- Vishnevsky, K. D. (1969a) Stanovlenie trekhslzhnykh razmerov v russkoj poezii. In: K. G. Petrosov (ed.). *Russkaya sovetskaya poeziya i stikhovedenie*. Moscow: s. n., pp. 207–217. (In Russian)
- Vishnevsky, K. D. (1969b) Vvedenie v stikhotvornuyu tekhniku XVIII veka. In: *Uchenye zapiski Ryazanskogo pedagogicheskogo instituta i Penzenskogo pedagogicheskogo instituta*. Vol. 81: *Voprosy stilya i zhanra v russkoj i zarubezhnoj literature*. Penza, pp. 3–16. (In Russian)

- Vishnevsky, K. D. (1985) Ekspressivnyj oreol pyatistopnogo khoreya. In: L. I. Timofeev (ed.). *Russkoe stikhoslozhenie: Traditsii i problemy razvitiya*. Moscow: Nauka Publ., pp. 94–113. (In Russian)
- Vostokov, A. Kh. (1817) *Opyt o russkom stikhoslozhenii*. 2nd ed. Saint Petersburg: Morskaya tipografiya Publ., [8], 1–141, 143–167, [8] pp. (In Russian)
- Zhirmunsky, V. M. (1975) *Teoriya stikha*. Leningrad: Sovetskij pisatel' Publ., 664 p. (In Russian)
-

Author:

Sergey I. Monakhov, ORCID: [0000-0002-0759-9998](https://orcid.org/0000-0002-0759-9998), e-mail: sergomon@gmail.com

For citation: Monakhov, S. I. (2019) One mechanism of Russian poetic language. *Journal of Applied Linguistics and Lexicography*, 1 (2): 315–330. DOI: [10.33910/2687-0215-2019-1-2-315-330](https://doi.org/10.33910/2687-0215-2019-1-2-315-330)

Received 4 August 2019; reviewed 28 August 2019; accepted 12 September 2019.

Copyright: © The Author (2019). Published by Herzen State Pedagogical University of Russia. Open access under CC BY-NC License 4.0.

ЦИФРОВЫЕ ПОТЕРИ В КОММУНИКАЦИИ ПРИ ПЕРЕХОДЕ ОТ «ОБЩЕНИЯ» К «ПЕРЕДАЧЕ ИНФОРМАЦИИ»

О. И. Северская✉¹

¹ Институт русского языка им. В. В. Виноградова РАН, 119019, Россия, г. Москва, ул. Волхонка, д. 18/2

DIGITAL COMMUNICATION LOSSES DURING THE TRANSITION FROM “COMMUNICATION” TO “INFORMATION TRANSFER”

O. I. Severskaya✉¹

¹ V. V. Vinogradov Russian Language Institute of the Russian Academy of Sciences, 18/2 Volkhonka Str., Moscow 119019, Russia

Аннотация. Статья посвящена проблемам коммуникации в цифровую эпоху. Цифровизация, преобразование информации в цифровую форму и одновременно внедрение современных технологий в любые сферы жизни общества противопоставляется информатизации, формированию оптимальных условий удовлетворения информационных потребностей пользователей за счет применения соответствующих технологий. Разница между ними в том, что в первом случае система может делать выбор за человека, а во втором она лишь помогает человеку, упрощая процесс освоения и применения культурных и языковых кодов. Одно из следствий цифровизации автор видит в изменении самого понятия коммуникации: это уже не «общение», а «обмен информацией», что зафиксировано в словарях лингвистических терминов.

Автор рассматривает инструментальность, преобладающую над коммуникативностью, как источник таких проблем, как гибридизация письменных и устных форм; сокращение и свертывание речи под влиянием текстинга; потеря смысла при аудировании; разрыв формы и содержания языкового знака; появление слов-стикеров, указывающих на объекты, действия и эмоции; выбор слов с меньшим набором сем в значении; замена словесных знаков и знаков препинания иконическими знаками (смайлами и эмодзи) и др. В качестве иллюстраций используются примеры текстов и диалогов, полученные методом включенного наблюдения, а также тексты из корпуса деловых писем. Анализ материала позволяет автору прийти к выводу: мозг современного человека работает как T9. При этом из всего репертуара языковых средств выбираются наиболее частотные, входящие в поверхностный слой активного лексикона, грамматикона и прагматикона говорящего и отвечающие доминирующему, «шлягерному» стандарту речи.

Ключевые слова: информация, сообщение, цифровая коммуникация, коммуникационные потери, информатизация, цифровизация.

Abstract. The article focuses on the issues of communication in the digital age. Digitalisation, which means converting information into digital form and simultaneously introducing modern technologies in all social areas, is contrasted with computerization — the process of forming the ultimate conditions to meet users' information needs by means of appropriate technologies. The difference between the two is that in the first case the system can make choices for an individual, while in the second it merely assists you by simplifying the process of mastering and using cultural and language codes. One of the consequences of digitalisation, in the author's opinion, is a change in the very concept of communication: it is no longer defined as “interpersonal communication”, but rather as “an exchange of information”, according to dictionaries of linguistic terms.

The author argues that instrumentality prevailing over communicativeness is the root of many issues, i.e. the merging of written and oral forms of language, the use of contracted and curtailed speech similar to texting, losing the plot while listening, broken links between the form and content of a language sign, the emergence of “sticker-words” indicating objects, actions, and emotions, choosing words with a smaller set of semes in their meaning, replacing words and punctuation with pictorial representations (emoticons and emoji), etc. As illustrations, the author uses samples of texts and dialogues obtained by means of applying the participant observation method, as well as texts from a corpus of business letters. Having analysed the material, the author concludes that the modern human brain works similarly to T9, selecting from the entire repertoire of linguistic means the most frequent ones which are included in the surface layer of the active lexicon, grammatical competence and pragmatics of the speaker and meet the dominant, “popular” standard of speech.

Keywords: information, message, digital communication, communication losses, computerisation, digitalization.

Цифровизация, под которой традиционно понимается переход на цифровой тип связи, явление двоякое: с одной стороны, сильно упрощается и ускоряется процесс коммуникации, с другой — изменяется сам этот процесс, причем не формально, а содержательно: если в конце XX в. *коммуникация* определялась как «общение», от греч. *κοινωνία* ‘делаю общим, связываю’ (Ярцева 1990, 233), то в начале XXI в. она становится «передачей информации», и в этимоне выделяются уже другие компоненты значения — ‘передаю, сообщаю’ (Азимов, Щукин 2009, 106). И потери в коммуникации при переходе от «общения» к «передаче информации» уже ощутимы. Кванты информации передаются, а вот связь и связность порой теряется. Между тем коммуникационологи всегда настаивали на различении информации и сообщения (Луман 1994, 32) и обращали внимание на то, что интересубъективная коммуникация «не может быть сведена к языковой передаче информации», а должна рассматриваться «одновременно с процессом достижения согласия» (Апель 2001, 52). Языку в коммуникации при этом отводится роль не только транслятора информации (Саакян 2016), но и средства достижения понимания.

Поколение Z, которое иногда еще называют «поколением большого пальца» (по имени главного «рабочего инструмента», облегчающего общение со смартфоном), и следующее за ним поколение α практически разучились говорить, не научившись толком писать: это поколения не текста, предполагающего связную и полную фиксацию человеческой мысли, а текстинга, процесса «подборки общепринятых сокращений, позволяющей в минимуме символов передать максимум информации» (Милеева, Кривоногова 2011, 65). Текстинг создает «письменную разговорность», и его влияние на речь проявляется в гибридизации форм устной и письменной речи.

Например, в автобусе можно услышать объявление: *Войдя на остановочном пункте, оплатите проезд. Оплата проезда производится путем поднесения средств оплаты к устройствам контроля и погашения*, полное характерных для письменного текста канцелярских оборотов и «тяжелых» синтаксических конструкций.

А по электронной почте приходят письма, содержащие разговорный эллипсис и написанные в смс-стиле: *В., подключила плз временно мне*, — при отсутствии знаков препинания не ясно: *подключила* — это утвердительный ответ на просьбу о подключении услуги? *подключила плз* — выраженная в грубой форме просьба (ср. подслушанное: *встала и закрыла, пожалуйста, окно!*) о временном предоставлении услуги? *плз временно мне* — просьба о временной адресации отчетов пишущему? или же это просьба о временном подключении самого автора письма к услуге?

Сами знаки препинания также испытывают воздействие текстинга и его инструментов. Так, почти полное исчезновение из переписки двоеточия и точки с запятой объясняется, во-первых, их меньшей доступностью на виртуальной клавиатуре (они требуют перехода на вторую ее страницу, где находятся также восклицательный и вопросительный знаки; на первой же странице доступны лишь точка и запятая), а во-вторых, тем, что знаки эти напоминают недорисованные смайлики «:)» и «;)». Точка и запятая нередко заменяются подходящими по смыслу эмоджонами, которые легко замещают и «эмоциональные» знаки препинания «?», «!» и «?!»; а скобки, напротив, начинают использоваться исключительно как знаки эмоциональности: «(((« и «)))».

Е. П. Савруцкая замечает также, что цифровизация «порождает новый тип коммуникативного поведения, все более регламентируемого символическими этикетными нормами и языково-лингвистическими элементами общения» (Савруцкая 2012, 133), а Л. Н. Мешкова, в свою очередь, указывает на диктуемое смс-стилем стремление к свертыванию речи, которое подразумевает не только пропуск знаков препинания, пробелов, гласных и исполь-

зование аббревиатур, но и «выбор среди синонимов слов с меньшим количеством букв» (Мешкова 2012, 50). Это приводит и к формированию новой знаковости, появлению своего рода смысловых и прагматически заряженных ярлычков.

Их можно назвать словесными знаками-стикерами. Стикером, как правило, называют наклейку или этикетку, предназначенную для уточнения информации об объекте, на которую стикер указывает. В сетевом сленге стикером называется «картинка», указывающая на объект, действие или эмоцию и фактически замещающая их. В нашем случае в роли таких указателей используются не смайлы или эмодзи, а слова. Как и бумажные стикеры, они *приклеиваются, липнут* к ситуациям и понятиям и *удерживаются* в речи.

Так, в речевой практике широко используется наряду с «шестисмайловым» набором знаков эмоций, если можно так выразиться, «пятисмайловый» набор маркеров включения говорящим режима вежливости, знаков коммуникативных намерений приветствия (*здравствуйте*), прощания (*до свидания*), просьбы (*пожалуйста*), благодарности (*спасибо*) и извинения (*извините*). Что не удивительно, так как пропагандируемое в сети «первое правило вежливости» гласит: «Вежливый человек не забывает произносить слова приветствия при встрече. Также он всегда прощается при расставании, извиняется, если причинил кому-то неудобства и благодарит за любые оказанные ему услуги» (Вежливый человек — какой он? 2016). Там же приводится и упомянутый выше краткий перечень универсально вежливых слов с комментарием: «Ежедневное употребление в своем лексиконе этих слов говорит о хороших манерах и высокой нравственности» (Вежливый человек — какой он? 2016). По-видимому, такого рода руководства и приводят к тому, что «волшебные» слова начинают доминировать и взаимодействовать с традиционными этикетными клише, то семантически дублируя их, то, напротив, вытесняя. В результате чего речь кажется либо слишком вежливой, либо невежливой, недотягивающей до общепринятого стандарта.

Классификационную роль играют и хештеги, превращающие полнозначные, эмоционально и оценочно нагруженные слова с ореолом коннотаций в ярлыки понятий или имена фреймов стандартных ситуаций, названных хештегированным словом или словосочетанием: *#ум, #честь, #совесть, #любовь, #яжемать* и под.

Все более заметен и выбор из ряда синонимов слов, обладающих не только меньшим числом знаков, но и меньшим набором сем в значении: появляются прономинализированные словесные стикеры, такие как *история* 'все, что угодно' или *шок* 'любая сильная эмоция', и слова, фактически являющиеся именами класса, например, *озвучить* (знак говорения, употребляемый вместо ряда глаголов: *произнести, сказать, довести до сведения, высказать вслух, проартикулировать, обнародовать, процитировать, упомянуть* и т. д.), *великий* (знак обладания положительно оцениваемым качеством, замещающий прилагательные *прославленный, знаменитый, известный, популярный, талантливый, способный* и т. п.).

Особый подкласс словесных стикеров образуют так называемые активные ссылки, или короткие ссылки в соцсетях на сети, сайты, профили пользователей, появляющиеся в текстах блогов и постов: они затрагивают и глубинные грамматические категории, такие как одушевленность-неодушевленность (активный линк превращает человека в бренд, знак самого себя), и синтаксис, усиливая рост аналитизма (ср. этикетку: *Пастила со вкусом ваниль*, рекламный текст: *Покупайте у нас в «Перекресток»!* и текст с активными ссылками: *Мы с Светлана Друговейко-Должанская, Мария Ровинская и Валерий Ефремов провели круглый стол на VI Педагогическом форуме в Сочи*).

Еще одно важное следствие цифровизации — модульный характер коммуникации (Савруцкая 2012, 133): общение сегодня идет в основном по шаблону, в котором обычно

прописаны и стандартные вопросно-ответные реакции. В условиях, когда письменная коммуникация превалирует над устной, участники «диалога» утрачивают способность к восприятию сообщений на слух и порой не «опознают» предмет речи, если говорящий указывает на него нешаблонно:

— Я хотел бы забронировать на вечер *стол* на троих часов на шесть...

— А сколько вас будет человек? —

или:

— Мне кофе *черный*, пожалуйста.

— Молоко, сливки *потребуется*? —

или же вовсе не слушают собеседника, следуя своей модульной программе (Северская, Селезнева 2019, 66–77):

— Будьте добры, пришлите ссылку на оплату путевки в Ялту, санаторий «Запорожье», одностельный номер, с 22 сентября на 7 ночей, с полупансионом.

— Да, конечно. Куда поедете? В Крым? В Сочи? Минеральные воды?

— В Ялту.

— Отель хотели бы три звезды? две? апартаменты?

— Санаторий «Запорожье».

— Без питания?

— С полупансионом.

— Когда поедете?

— 22-го, на 7 ночей.

— А сколько вас будет?

Отсутствие привычки к аудированию приводит и к смешению слов-паронимов: некоторые путают такие «категории состояния», как *смеркалось* и *сморкалось*, блуждают в *дербях* вместо *дебрей*, *пишут эпизодические жалобы* вместо *апелляций* (Северская 2018a; 2018b). Современный человек ориентируется на «слуховую память», выбирая слова из своего активного словаря и не слишком заботясь о смысле. Мозг при этом действует как T9 — предикативная система набора текстов для смартфонов и планшетов (от *Text on 9 keys* «текст на 9 кнопках»): T9, используя встроенный словарь, пытается предугадать, какое слово имелось в виду, и наиболее часто употребляемые слова подставляются первыми. Результат — ошибки типа *Олень*, *Вы сможете прийти на собеседование в 13 часов?* или *Не ставь машину около дома, там дед падает с крыши постоянно* (((, которые мало чем отличаются от замены *симпатичной светлой палитры матирующей пудры* на *поллитру*, — замены, произведенной не гаджетом, а человеком. К сожалению, это закономерно: ссылаясь на концепцию Ю. Хабермаса, различавшего в социальных действиях инструментальные

и коммуникативные (Хабермас 2000, 199), Савруцкая констатирует (Савруцкая 2012, 132) актуальное преобладание инструментальной составляющей в современной цифровой коммуникации.

Если в эпоху информатизации компьютерные технологии облегчали, упрощали и ускоряли процесс освоения культурных кодов, и это мало чем отличалось от усвоения информации из книг, то цифровизация, по мнению О. Липиной, таит в себе гораздо больше опасностей, поскольку цифровая система «может действовать независимо и обладает аналитическими и прогностическими функциями, иными словами, она может делать выбор за человека» (Липина 2018): «Называть ее искусственным интеллектом нет полного основания. Скорее всего, в настоящий период перед нами средний вариант: уже не мышление человека, но еще не сознание машины» (Липина 2018). Таким образом, остро стоит задача, оценив риски потерь при переходе от «общения» к «передаче информации», использовать возможности «цифры» для возвращения коммуникации ее человеческого измерения.

Литература

- Азимов, Э. Г., Щукин, А. Н. (2009) *Новый словарь методических терминов и понятий (теория и практика обучения языкам)*. М.: ИКАР, 446 с.
- Апель, К.-О. (2001) Априори коммуникативного сообщества и основания этики. В кн.: К.-О. Апель. *Трансформация философии*. М.: Логос, с. 263–335.
- Вежливый человек — какой он? Качества вежливого человека. человека. (2016) *ФБ.ру*, 25 марта. [Электронный ресурс]. URL: <http://fb.ru/article/238132/vejliviy-chelovek---kakoy-on-kachestva-vejlivogo-cheloveka> (дата обращения 31.07.19).
- Липина, О. (2018) Чем отличается информатизация от цифровизации? *Ру.Ка*, 17 августа. [Электронный ресурс]. URL: <http://ruzaknary.ru/2018/08/17> (дата обращения 31.07.19).
- Луман, Н. (1994) Понятие общества. В кн.: А. О. Бороноев (ред.). *Проблемы теоретической социологии*. СПб.: Петрополис, с. 25–42.
- Мешкова, Л. Н. (2012) Текстинг как явление современной культуры. *Известия Пензенского государственного педагогического университета им. В. Г. Белинского*, 27: 49–53.
- Милеева, М. Н., Кривоногова, О. А. (2011) Текстинг в условиях мобильной коммуникации. *Известия высших учебных заведений. Серия «Гуманитарные науки»*, 2 (1): 65–68.
- Саакян, Л. Н. (2016) Новые медиа: язык СМИ и социальных сетей. *Русский язык за рубежом*, 4 (257): 93–99.
- Савруцкая, Е. П. (2012) Проблемы коммуникации в контексте социокультурной реальности информационного общества. *Вестник Нижегородского университета им. Н. И. Лобачевского*, 1 (3): 132–137.
- Северская, О. И. (2018a) Каким должен быть школьный словарь паронимов? В кн.: *Динамика языковых и культурных процессов в современной России*. Вып. 6. СПб.: Общество преподавателей русского языка и литературы, с. 506–511.
- Северская, О. И. (2018b) Орфографические нормы сквозь призму паронимии и сходнозвучия. В кн.: Л. А. Вербицкая, С. И. Богданов, С. В. Друговейко-Должанская и др. (ред.). *Языковая норма. Виды и проблемы: материалы V международного педагогического форума (Сочи, Россия, 3–4 декабря 2018 г.)*. СПб.: Изд-во РГПУ им. А. И. Герцена, с. 191–200.
- Северская, О. И., Селезнева, Л. В. (2019) *Эффективная бизнес-коммуникация. «Вошебные таблетки» для деловых людей*. М.: Эксмо, 410 с.
- Хабермас, Ю. (2000) *Моральное сознание и коммуникативное действие*. СПб.: Наука, 377 с.
- Ярцева, В. Н. (ред.). (1990) *Лингвистический энциклопедический словарь*. М.: Советская энциклопедия, 685 с.

References

- Apel, K.-O. (2001) Apriori kommunikativного сообщshchestva i osnovaniya etiki. In.: K.-O. Apel. *Transformation der Philosophie [Transformation of philosophy]*. Moscow: Logos Publ., pp. 263–335. (In Russian)
- Azimov, E. G., Shchukin, A. N. (2009) *Novyy slovar' metodicheskikh terminov i ponyatij (teoriya i praktika obucheniya yazykam)*. Moscow: IKAR Publ., 446 p. (In Russian)

- Habermas, Jü. (2000) *Moralbewusstsein und Kommunikatives Handeln*. Saint Petersburg: Nauka Publ., 377 p. (In Russian)
- Lipina, O. (2018) Chem otlichaetsya informatizatsiya ot tsifrovizatsii? *Ru.Ka*, 17 August. [Online]. Available at: <http://ryazankray.ru/2018/08/17> (accessed 31.07.19). (In Russian)
- Luhmann, N. (1994) Ponyatie obshchestva. In.: A. O. Boronoev (ed.). *Problemy teoreticheskoy sotsiologii*. Saint Petersburg: Petropolis Publ., pp. 25–42. (In Russian)
- Meshkova, L. N. (2012) Teksting kak yavlenie sovremennoj kul'tury [Texting as a phenomenon of contemporary culture]. *Izvestiya Penzenskogo gosudarstvennogo pedagogicheskogo universiteta im. V. G. Belinskogo*, 27: 49–53. (In Russian)
- Mileeva, M. N., Krivonogova, O. A. (2011) Teksting v usloviyakh mobil'noj kommunikatsii. *Izvestiya vysshikh uchebnykh zavedenij. Seriya "Gumanitarnye nauki"*, 2 (1): 65–68. (In Russian)
- Saakyan, L. N. (2016) Novye media: yazyk SMI i sotsial'nykh setej [New media: The language of the mass media and social networks]. *Russkij yazyk za rubezhom — Russian Language Abroad*, 4 (257): 93–99. (In Russian)
- Savrutskaya, E. P. (2012) Problemy kommunikatsii v kontekste sotsiokul'turnoj real'nosti informatsionnogo obshchestva [The problems of communication in the context of sociocultural reality of the information society]. *Vestnik Nizhegorodskogo universiteta im. N. I. Lobachevskogo — Vestnik of Lobachevsky University of Nizhni Novgorod*, 1-3: 132–137. (In Russian)
- Severskaya, O. I. (2018a) Kakim dolzhen byt' shkol'nyj slovar' paronimov? In: *Dinamika yazykovykh i kul'turnykh protsessov v sovremennoj Rossii*. Vol. 6. Saint Petersburg: Obshchestvo prepodavatelej russkogo yazyka i literatury Publ., pp. 506–511. (In Russian)
- Severskaya, O. I. (2018b) Orfograficheskie normy skvoz' prizmu paronimii i skhodnozvuchiya [Spelling norms through the prism of paronymy and euphony]. In: L. A. Verbitskaya, S. I. Bogdanov, S. V. Drugovejko-Dolzanskaya et al. (ed.). *Yazykovaya norma. Vidy i problemy: Materialy V mezhdunarodnogo pedagogicheskogo foruma (Sochi, Russia, 3–4 December 2018.)*. Saint Petersburg: Herzen State Pedagogical University of Russia Publ., pp. 191–200. (In Russian)
- Severskaya, O. I., Selezneva, L. V. (2019) *Effektivnaya biznes-kommunikatsiya. "Volshebnye tabletki" dlya delovykh lyudej*. Moscow: Eksmo Publ., 410 p. (In Russian)
- Vezhlivyj chelovek — kakoj on? Kachestva vezhlivogo cheloveka. *FB.ru*, 25 March. [Online]. Available at: <http://fb.ru/article/238132/vejlivyyi-chelovek---kakoy-on-kachestva-vejlivogo-cheloveka> (accessed 31.07.19).
- Yartseva, V. N. (ed.). (1990) *Lingvisticheskij entsiklopedicheskij slovar'*. Moscow: Sovetskaya entsiklopediya Publ., 685 p. (In Russian)

Сведения об авторе:

Ольга Игоревна Северская, ORCID: 0000-0002-6277-9756, e-mail: oseverskaya@yandex.ru

Для цитирования: Северская, О. И. (2019) Цифровые потери в коммуникации при переходе от «общения» к «передаче информации». *Journal of Applied Linguistics and Lexicography*, 1 (2): 331–336. DOI: 10.33910/2687-0215-2019-1-2-331-336

Получена 14 августа 2019; прошла рецензирование 28 августа 2019; принята 12 сентября 2019.

Права: © Автор (2019). Опубликовано Российским государственным педагогическим университетом им. А. И. Герцена. Открытый доступ на условиях лицензии CC BY-NC 4.0.

Author:

Olga I. Severskaya, ORCID: 0000-0002-6277-9756, e-mail: oseverskaya@yandex.ru

For citation: Severskaya, O. I. (2019) Digital communication loss in the transition from “communication” to “information transfer”. *Journal of Applied Linguistics and Lexicography*, 1 (2): 331–336. DOI: 10.33910/2687-0215-2019-1-2-331-336

Received 14 August 2019; reviewed 28 August 2019; accepted 12 September 2019.

Copyright: © The Author (2019). Published by Herzen State Pedagogical University of Russia. Open access under CC BY-NC License 4.0.

THE PROJECT OF A DEEPLY TAGGED PARALLEL CORPUS OF MIDDLE RUSSIAN TRANSLATIONS FROM LATIN

E. G. Sokolov^{✉1}

¹ Institute for Linguistic Studies, Russian Academy of Sciences, 9 Tuchkov Ln., Saint Petersburg 199053, Russia

Abstract. Tagged parallel corpora are powerful tools for the analysis of natural language. Moreover, for historical linguistics, whose most peculiar shortcoming is lack of living native speakers, corpora — as paper or electronic collections of written texts — are the main source of linguistic information. Old and Middle Russian are well-documented languages, and a host of manuscripts in both idioms — including those containing numerous translations — are available for investigation. Nevertheless, up to now there is no parallel translational corpus of Middle Russian. Thus, a number of important written sources containing information valuable for linguists, literary scholars and historians cannot be studied properly. This article provides a preliminary account of the project of a deeply tagged parallel corpus of Middle Russian translations from Latin. Such corpus may prove useful in the formal description of the translation techniques of the time, which may help with dividing the anonymous texts of the time into several groups based on their language features. Such grouping may help with authorship attribution and, consequently, with incorporating each translation into a proper cultural landscape.

From the linguistic point of view, such corpus could provide researchers with crucial information on the vocabulary, morphology and syntax of Middle Russian with an emphasis on the argument structure of the verbs, usage of borrowed lexical items and set expressions and professional skills of the ancient translators. The article gives an outline of the crucial features of the prospective Middle Russian translational corpus, its possible primary contents, text standardization and annotation principles, as well as the reasons for not using a theory-neutral syntactic apparatus, characteristic of the existing historical corpora of ancient Indo-European languages, such as TOROT or PROIEL. An explanation of how the potential users of this corpus could benefit from our non-standard tagging principles is given.

Keywords: Middle Russian, Church Slavonic, Latin, translation, electronic corpora, syntactic alignment.

Introduction

Short description of the project

A parallel corpus includes a number of texts in one language with their translation into another; the corresponding fragments of the original and the translation are aligned. This usually provides tools that assist in finding not only lexical information, but other linguistic data, e. g., morphological or syntactic. Although there exist a number of Old and Middle Russian corpora (Mitrenina 2014), no parallel corpus of Old or Middle Russian translations was ever developed.

This article presents the project of deeply tagged parallel corpus of Middle Russian translations from Latin, i. e. Latin — Russian¹ translational corpus with parallel syntactic alignment (hereinafter LRC), containing the Russian translations from Latin made between the end of the 15th century and the beginning of the 17th century — that is, roughly, within period of a hundred years. It will provide a wide range of researchers with a modern instrument for an in-depth analysis of a significant layer of premodern Russian culture, namely the pretheoretic translational activity in the pre-Petrine Russia. The tasks of the project include:

¹ The notion “(Middle) Russian” here and further refers rather to the origin of the translations than to their language, comprising both Middle Russian and Church Slavonic text translated from Latin.

- (1) formulating the general principles for syntactic, semantic, morphological and lexical annotation of the texts respectively in Latin and Russian parts of LRC;
- (2) formulating the alignment principles for text pairs in LRC, especially the principles of syntactic alignment within such pairs;
- (4) formulating the main principles of a bilingual glossary built using the LRC alignment;
- (5) developing and adjusting the electronic tools for implementation of the objectives (1–4);
- (7) formulating the guidelines for further annotation of LRC and other similar corpora in future.

The fulfillment of these tasks will let us launch the first deep annotated parallel historical corpus of Russian language, which will substantially enhance the research abilities of the scholars concerned with history of Slavic languages and cultures.

Some general reasons for creating LRC

Translations (mostly from Greek and Latin) constitute a substantial part of the written sources of Old and Middle Russian origin, and are undoubtedly of considerable importance for both linguists and literary scholars. These translations can provide researchers with important data concerning historical grammar and history of Russian language, including such topics as lexical, idiomatic and syntactic borrowings and their impact on written language, as well as translation principles, their variation and gradual changes. Russian pre-Petrine translations could be also of some interest for the researchers concerned with contrastive grammar description or language typology, because the process of aligning bilingual corpora provides valuable information about both the source and target languages and their grammatical properties (Grishman 1999, 225). They are also a great source of cultural and historical information, and thus must be the matter of interest for historians and literary scholars, who may also require some additional linguistic information in order not to mistake in ascribing the text to a wrong author.

For historical linguists studying the premodern languages whose only sources are sets of written texts, such sets — that is in fact, corpora (though not obligatorily in electronic format) — are the only reliable sources of linguistic data. To make such data verifiable, reliable and as exhaustive as possible (which is necessary for a serious study of any subject) they have to be easily collectable and extractable. This can only be done by means of an electronic corpus. So there exists an urgent need for a translational corpus of Old and/or Middle Russian.

To make the extraction of the information mentioned above possible, the corpus must be able to give a precise account of the structural relationships between the translation and its original. This means that the translational corpus must contain not only translations themselves, but also their originals paired to them, and that such pairs (which we will hereinafter call *parallel texts* or *bitexts*) have to bear the metalinguistic information mapping the units of the original into the units of its translation, i. e., to be aligned. The simplest way of aligning parallel texts is word alignment. But the alignment at the word level poses serious difficulties due to the complexity of the correspondences across the languages (Grishman 1999, 226). There is in fact next to no word-to-word correspondence even in a very literal translation. That is why it seems much more reasonable to compare the syntactic units of the parallel texts, not their lexical units (Grishman 1999, 226), proceeding top-down, that is, from the higher grammar units to the lower ones.

There is another reason for providing the translational corpus with parallel syntactic alignment, which is especially important for a corpus consisting of texts preserved in manuscript tradition. In the process of multiple copying a handwritten text undergoes multiple changes at nearly all levels, and syntax is the only feature of the handwritten text which remains relatively stable for a long period of time (Tomelleri 2011, 219), so it is desirable that the parallel texts in a historical corpus have a syntactic alignment.

Lack of translational corpora for pre-Petrine Russian texts

As of now there exist a substantial number of electronic resources containing historical corpora of Slavic languages, for example, the historical and Church Slavonic subcorpora of the Russian National Corpus, the MANUSCRIPT project, the SKAT project, the OCS subcorpus of the PROIEL project, the TOROT project, and some others. As a rule, such corpora are provided with lexical and morphological annotation, but most of them lack any tools for syntactic analysis of the texts. The only exceptions are PROIEL, TOROT and SKAT. The first two projects were initially designed to be syntactic treebanks (Haug et al. 2009; Eckhoff, Berdičevskis 2016, 63), i. e. sets of texts with syntactic trees associated with their units, while the syntactic module of the last project is still being developed (Alekseeva 2014).

So, in fact there are only two corpora containing the syntactic representation of OCS, Old and Middle Russian data, and both of them cannot meet the requirements imposed by our goal. The PROIEL corpus contains only three OCS codices (*Marianus*, *Suprasliensis* and *Zographensis*) and three Old Russian texts (*Codex Laurentianus*, *The Taking of Pskov* and *The Tale of Luka Kolocskij*) and thus is of nearly no interest to us; the TOROT project offers a lot of Old and Middle Russian sources, but doesn't aim at translation studies, containing mostly original texts.

That is why we believe that there is an urgent need for a linguistic project like LRC, focusing on the syntactic and semantic representation of aligned bitexts.

Prospective features of the LRC project

As we have already mentioned, the LRC project must be designed to meet the requirements of historical translation studies. First of all, to study a translation is to describe and explain its technique, i. e. reveal the transfer rules underlying the process of rendering a text in the source language into a text in the target language. From this point of view alignment is pairing of a subset of the nodes in the source syntactic tree with a subset of the nodes in the target syntactic tree (Grishman 1999, 228). But each set of nodes is, in fact, a particular manifestation of an abstract grammar construction; for instance, the pair *stante illa domo ~ стоящу дому* (Fedorova 1999b, 98) is a manifestation of the *ablativus absolutus* in its Latin part and of the *dativus absolutus* in its Church Slavonic part, thus a manifestation of the *ablativus absolutus ~ dativus absolutus* correspondence between the source and goal languages. Each text, Latin or Russian, must be viewed as a set of units corresponding to a certain set of grammatical constructions. Hence, the syntactic alignment of the LRC bitexts must let the researcher establish the correspondence between the set of grammatical constructions in a source language and the set of grammatical constructions in the respective target language.

Any text can be represented by a finite sequence, or *string*, of word forms, which can be divided into several *substrings* (Partee, ter Meulen, Wall 1990, 433–435).

Furthermore, any text can be represented by a finite ordered set of separate *syntactic trees*. Syntactic tree is a tree in graph theoretic sense, over which a certain *relation* is defined. If a tree depicts the *part – whole relation* between the substrings of a certain string, it is called a *phrase structure tree* or a *phrase marker* (hereinafter PST); if a tree represents the *dependency relation* which holds between its single nodes (Gaifman 1965, 306), it is called a *dependency tree* (hereinafter DT). There is the third relation, *dominance*, defined as follows: a node α is said to *dominate* a node β , if a connected sequence of branches can be extended from α to β (Partee, ter Meulen, Wall 1990, 433–435). It is obvious that dominance holds both between the nodes of PST and DT.

Alignment can be defined as pairing of a subset of the nodes in the source syntactic tree with a subset of the nodes in the target syntactic tree (Grishman 1999, 228), see (fig. 1).

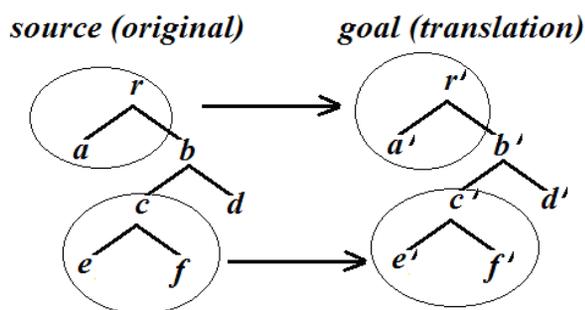


Fig. 1. Aligning the corresponding trees in the source and the goal texts

Syntactic construction. As any syntactic tree represents certain structural configuration of its non-terminal nodes, and this structural configuration may be preserved even if the word forms in the terminal nodes are replaced by other forms of the same distributional class, we may call such stable structural relation *a syntactic construction*. For instance, Latin utterances (a) and (b)² both contain subordinate clauses whose subject noun phrase bears accusative case and verb takes the infinitive form. They are examples of a specific construction called *accusativus cum infinitivo*, which has the following features: (1) it has the form of a clause; (2) it is the dependent of a verb or participle (here *tradidit* and *compertum*); (3) its subject bears accusative case; (4) its predicate is represented by an infinitive form of a verb.

(a) [*tradidit Herodotus*] ... *cynamomum in auium nidis reperiri* ‘

[Herodotus tells ...] that cinnamon can be found inside bird’s nests’

(b) [*compertum*] ... *cynamomum longissime ab omni Aethiopia gigni*

‘[is revealed] ... that cinnamon grows as far as possible from any land associated with Ethiopia’

Let us assume that each particular source text is generated by means of the specific set of syntactic constructions $C := \{c_1, c_2, c_3 \dots c_n\}$, and that there is the set of grammar constructions $K := \{k_1, k_2, k_3 \dots k_n\}$ which corresponds to it in the target language and enables the generation of the target text. Now it is possible to introduce the notion of *translation technique*.

Translation technique. Let A be a source text and A' be its translation. Then the translation technique M for the pair $\langle A, A' \rangle$ is the set of all the pairs of *syntactic constructions* or *syntactic subtrees* $\langle \delta, \delta' \rangle$ ordered by a binary relation R defined as follows: $\delta R \delta' := \delta$ is translated via δ' .

$$M := \{ \langle \delta, \delta' \rangle \mid (\delta \in A) \ \& \ (\delta' \in A') \ \& \ (R \langle \delta, \delta' \rangle = 1) \}$$

Hence the syntactic alignment of the LRC bitexts must let the researcher establish the correspondence between the set of grammatical constructions in a source language and the set of grammatical constructions in the respective target language in order to define the translation technique for a certain bitext, like that given below (fig. 2) for “The Letter on the Moluccas” written by *Maximilianus Transsylvanus*.

The translation technique can seriously differ for different bitexts, which means that for two different translation techniques T_1 and T_2 there must be two correspondingly different sets of relations between the source and target constructions, say, R_1 and R_2 . The LRC has to be able

² Here and further Latin and Russian examples are taken from The Letter on the Moluccas originally written in Latin by Maximilianus Transylvanus and translated into Russian in the mid-1520s. We have been studying this translation and its Latin source since 2013 and thus will often give examples from this text in the following sections of our article.

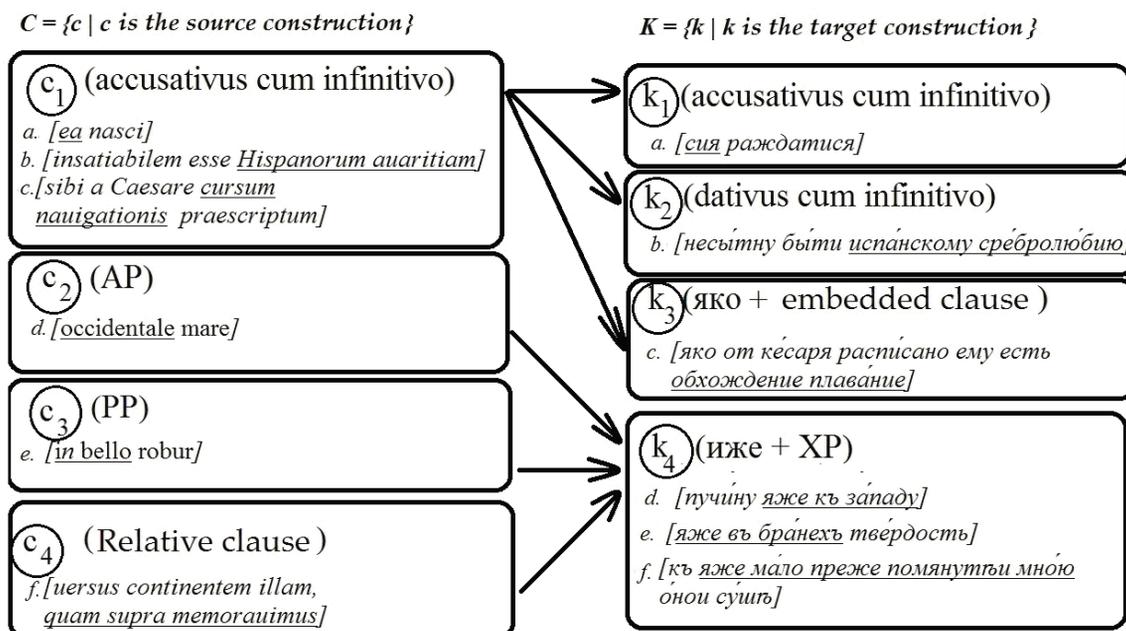


Fig. 2. An informally represented fragment of translation technique for “The Letter on the Moluccas”

to give the researcher a convenient opportunity for rendering the manually aligned pairs of syntactic nodes of a bitext into pairs of grammatical constructions (whereby the frequency of each construction must be taken into consideration as well), resulting in the set of transfer rules for each bitext. Having such sets of transfer rules, a scholar will be able to compare the translational techniques of different texts and to draw verifiable conclusions about the degree of propinquity of these texts.

For example, let us denote the bitext of “The Letter on the Moluccas” as A and imagine that there is a certain bitext B in which $R := \{(c_1, k_1), (c_1, k_3), (c_3, k_4)\}$, that is *accusativus cum infinitivo* is always translated either with *accusativus cum infinitivo* or with an embedded clause introduced by a complementizer *яко*, and the Church Slavonic non-finite construction with *иже*, *еже*, *яже* is the translation only for a prepositional phrase. In this case the machine will easily draw a Venn diagram (fig. 3) and let us see that $R(B) \subseteq R(A)$:

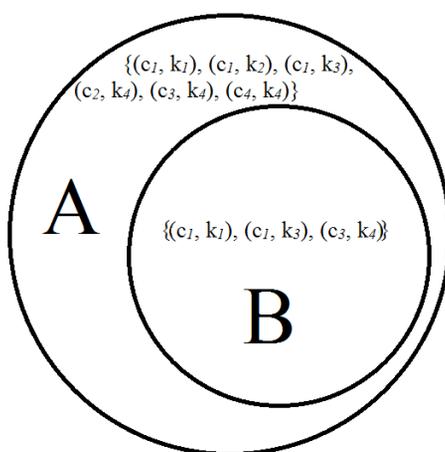


Fig. 3. An example of the Venn diagram

If all the bitexts in LRC have their own sets R , then for each n bitexts it will be possible to draw a Venn diagram consisting of n parts, each corresponding to a particular set R . If each pair in each set R is provided with its frequency rate, it will be possible to compare the frequency of a certain construction throughout all the bitexts available in the corpus. Moreover, it will enable the automatic assessment of propinquity degree for each n bitexts in corpus, preventing various scholars (especially non-linguists) from attribution faults caused by arbitrary and inconsistent reasoning.

Let us consider an example of such doubtful attribution. The first editors of the Russian version of “The Letter on the Moluccas”, N. A. Kazakova and L. G. Katuškina, ascribe this translation to Dimitri Gerasimov (Kazakova, Katuškina 1968, 237–238). Their reasoning runs as follows. Dimitri was one of the most prominent translators of the time; besides, he was well-known for his diplomatic activities and took part in the famous embassy to Rome, where he could have bought a Latin exemplar of the book mentioned above; finally, Mikhail Medovartsev, the scribe who prepared the only known copy of the text, was familiar to Dimitri. Basing entirely on these extralinguistic assumptions, the editors proceed from historical facts to linguistic conclusions about the structure of the text itself, noting that the translation preserves the literal manner typical for Gerasimov and follows the syntax of the Latin source (Kazakova, Katuškina 1968, 234). These conclusions were later repeated in Kazakova’s monograph (Kazakova 1980) and D. O. Tsytkin’s article (Tsytkin 1990). Nevertheless, in 1990 there appeared an article written by a German scholar Elke Wimmer (Wimmer 1990), which convincingly proved that Dimitri had no hand in this translation, giving strong linguistic reasoning by the expedient of comparison of the extracts from “The Letter on the Moluccas” with another translation surely made by Gerasimov, which demonstrated that the translation technique of The Letter differed considerably from what one could expect from a Gerasimov’s translation.³ The same conclusions were drawn in (Sokolov 2014), which is also based on some linguistic observations, despite the fact that the author was not familiar with Wimmer’s article at that time. This example proves the necessity of linguistic studies for correct attribution of ancient translated texts, and the more precise such studies are, the more reliable will the attribution be. That is why we believe that the propinquity assessment mechanism similar to the one described in the initial part of this section must be implemented within the LRC project.

Primary set of texts

This section presents the primary set of texts which could become the basis of the project. Some of them have been already published, these are the translation of William Durandus’ “Rationale divinatorum officiorum” (Durandus 2012), the translation of Nicolaus de Lyra’s “Probatio adventus Christi” (Fedorova 1999a), the so-called “Pravila gramatichnye” (Tomelleri 1999), some parts of the so-called Bruno’s Psalter (Tomelleri 2004), Maxim Grek’s translation of Piccolomini’s “De Captione urbis Constantinopolitanae” (Kloss 1975, 55–61; the text itself: Kloss 1975, 59–61), Guido de Columna’s “Historia destructionis Troiae” (Tvorogov 1972), the translation of Pomponius Mela’s “Chorographiae liber” (Matasova 2014). There are also several texts which need a re-edition, because their primary edition was weak from many points of view, e. g. the translation of Transsylvanus’ “De Moluccis insulis... epistola” (“The Letter on the Moluccas”) (Kazakova, Katuškina 1968). Some texts have never been published but are also of great importance for us, for instance, the so-called “Book of Saint Augustine” (Kalugin 2001). All these texts must be included into LRC as its basic component.

³ See also (Wimmer 2005, 74).

Thematically related projects and software tools they use

Several projects might shed some light on how to implement the LRC project avoiding the difficulties which have already been overcome by other scholars.

An electronic database for Russian and Church Slavonic handwritten sources developed in the Vinogradov Institute of Russian Language allows for manual and semi-automatic markup, as well as automatic formation of lexical and grammatical indices, and is provided with a GUI which makes its application more user-friendly (Arkhangelskiy, Mishina, Pichkhadz 2014, 102). More importantly, its structure, based on the YAML files, allows for marking up strings of words and syntactic phrases (Arkhangelskiy, Mishina, Pichkhadz 2014, 102–103). One of the most striking features of the project is the ability to ascribe certain characteristics not to a single token but to a unit as a whole. This system also lets the annotator establish the correspondences between units within such parallel texts where units from both sides may be of arbitrary length (Arkhangelskiy, Mishina, Pichkhadz 2014, 103).

There also exists a morphological tagger suited specifically for processing Middle Russian texts, the so-called RNC analyzer, developed at Higher School of Economics (Moscow) for annotating the Middle Russian subcorpus of the Russian National Corpus (Berdičevskis, Eckhoff, Gavrilova 2016, section 1). The RNC analyzer is a rule-based system in the UniParser format (see below), which is able to give a grammatical annotation to any text for whose language there exists a properly formed grammar description. It could be also useful as an example of a successful automatic tagger for Middle Russian (even though we do not think there is an urgent need for fully automatic annotation in a relatively small corpus like LRC). More on the principles of the UniParser-based RNC analyzer one may find in (Gavrilova, Shalганova, Liashevskaja 2016).

Basic technical details

LRC must provide online access to the electronic editions of Latin and Middle Russian texts, as well as a set of search tools for processing the data contained in these texts. The texts would be stored as XML documents for the purposes of their accessibility and easy processing. As to the search and annotation instruments, some considerations on that topic will be given in the following parts of the current article.

Preprocessing and annotation levels

In the next parts of the article we will give a brief outline of the steps for processing and annotating the texts for LRC. First of all, let us shortly list these steps in the following scheme (fig. 4).

This is the tentative order in which the required steps will be applied to the texts of the corpus. Undoubtedly it could be much more fine-grained and detailed, but this is only a preliminary outline, so we kindly ask the reader to forgive us for such a short description of the annotation process. We hope that some details of the process will be further clarified below.

Preprocessing and normalizing the text

Handwritten texts may considerably differ as to their graphics, orthography and punctuation systems. If one wants to establish a searchable and uniform corpus of such texts, one cannot let the chaotic richness of particular writing systems belonging to various scribes or scribal traditions survive the digitalization of the text. Thus, every text chosen for LRC has to be previously normalized in respect to its graphics, orthography and punctuation systems.

In the following subsection we will try to give a brief outline of the normalization system which we tend to develop for the Corpus.

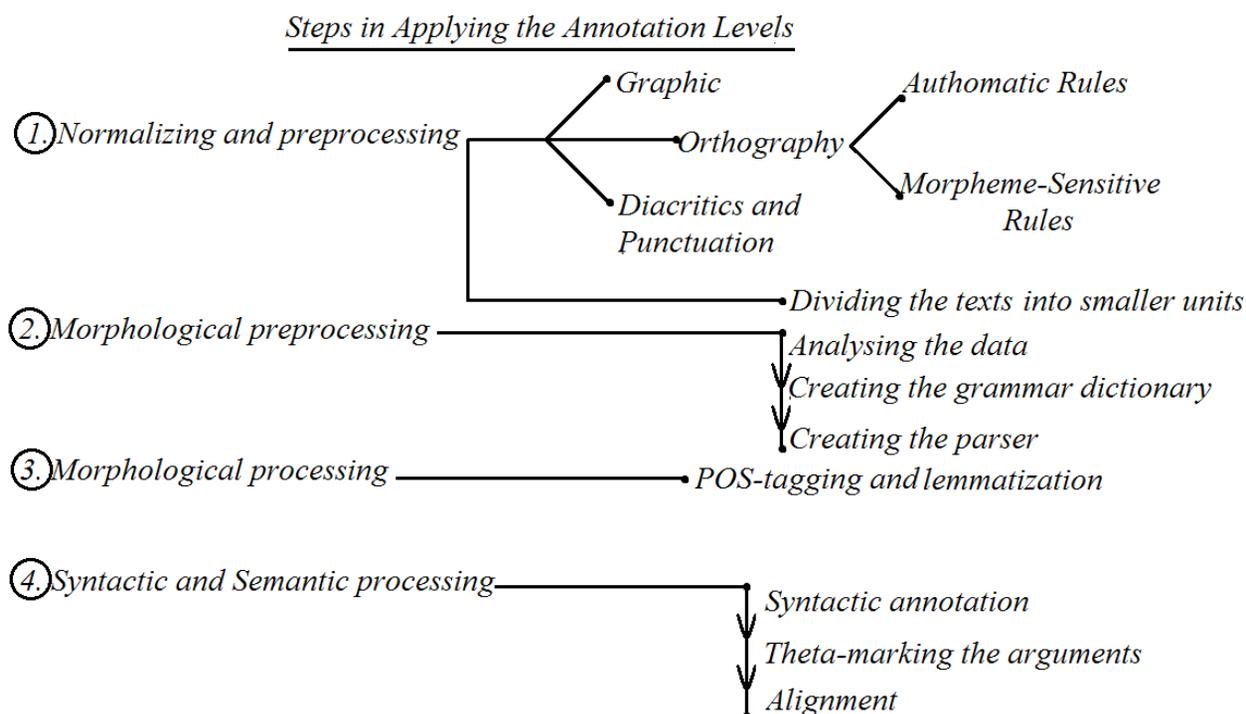


Fig. 4. Steps in applying the annotation levels in LRC

Graphics

In order to give a unified representation of the handwritten sources in LRC, the graphics system used in LRC must fit some general criteria, which are given below.

Preserving the Necessary Graphemes. The first criterion demands that the texts have to preserve as many graphic units from the original manuscript copies as necessary, no less and no more. In general, if there is a distinction between the graphic unit G_1 and the graphic unit G_2 , this distinction must be preserved, if it regularly conveys some necessary grammatical information and is typical for the text (or texts). For example, there are two graphic units ϵ and \mathfrak{B} . We can see that the second one, \mathfrak{B} , is regularly used in a certain set of morphemes, where ϵ is never found, e. g. in certain roots (*крѣн-*: *крѣн-каа*, *у-крѣн-лѣше*) or in the * \bar{a} -declension and * \bar{o} -declension allomorphs of the locative singular morpheme: *по вѣноу пучѣн-ѣ* and *вѣмьст-ѣ*. Correspondingly, if there is no morphemic or other grammatical distinction inside a pair of graphic units (that is, if they are both possible in the same morpheme under the same circumstances), they are in fact free graphic variants of the same grapheme, and one of such graphic units must be eliminated and substituted by the other one everywhere, like in the following examples:

(a) \mathcal{A}' : \mathcal{A}'

There are some instances where the use of \mathcal{A}' or \mathcal{A}' is fully arbitrary, as in the following pair of contexts, where both units occur in the same position and in the same word:

великую пучину, и языки синейскыа : страну вѣну и Азыкъ
sinum magnum & Sinarum populos : regionem et gentem

It is obvious that in the examples given above (opposite to the situation in modern Church Slavonic) there is no real distinction between the first and the second grapheme.

(b) *y : r : oy : ov*

The examples given below demonstrate that the use of these four graphic units doesn't exhibit any grammatic or semantic differences, or complementary distribution between them:

ура(з)умѣвъ : оудобнѣ : убо : ѡтудоу : не онъ : онбо : трѣбовю(т)

That is why the four variants have to be replaced by only one of them, namely the variant *y*.

Finally, some pairs of graphic units are superfluous from the point of view of our knowledge about the phonological system of the language of the time. For instance, even though the graphic unit *s* in the pair *s : z* has a specific distribution, being used in a certain set of roots, e. g. *сѣпѣ-, сум-, съл-*, in the 16th century it neither reflects any special phoneme distinct from /z/ nor is used to distinguish homophonous morphemes. That is why it must also be replaced by its more common counterpart, *z*.

Avoiding the allographic variation. Once the set of necessary graphic units, i. e. graphemes, is defined, and all alternations in the same position are eliminated, we must also eliminate the sets of variants whose members differ only with respect to their mutual complementary distribution. If there is any complementary distribution (i. e. allographic variation) within a grapheme, the number of allographs must be reduced to one. Below we will consider the most widespread example, the pair *ï : u*.

From the very time of the so-called Second South Slavic Influence (Grot 1894, 59), the pair of graphic units *ï : u*, whose distribution had been rather free before, became allographs of the grapheme *u*, where *u* served as a primary allograph, and the second allograph, *ï*, was placed before the vowels (this rule stayed nearly untouched⁴ until the 1918 orthographic reform).

Undoubtedly, there is no need to preserve such allographic alternations in LRC, so all such cases are subject to reduction. The discussed case, for example, has to be reduced to a single graphic variant of the *ï : u* grapheme, namely *u*. For instance, in both forms *сѣмныа* and *помолѣша* the pair *ï : u* should receive the same representation as *u*: *зѣмныа* and *помолѣша*.

Reducing unnecessary diacritics. The Second South Slavic Influence reintroduced a number of long-forgotten and unnecessary diacritic signs borrowed from the Greek script. Most of them, like aspiration signs, are *completely useless*. Let us call them the signs of the *first type*. The other ones (which we will call *the signs of the second type*), usually stress signs, *may be of some use* for us, primarily because they indicate the place of the word stress. The signs of the first type must be completely eliminated from the text. Among the signs of the second type, all signs that have similar function must be reduced to one. After such procedures the text will preserve the stress signs, but lose all the unnecessary South Slavic ornamental elements, like in the following example:

но сѣмныа сълности . и страны о́ноа лю́тости убо́явшеса , помолѣша своего корабленачалника маггелана → но зѣмныа зьлности и страны о́ноа лю́тости убо́явшеса, помолѣша своего корабленачалника Маггелана.

In this way all the linguistic information is preserved, while the text makes one more step towards normalization.

⁴ Except the cases of morphemic borders like шести-аршинный, пяти-этажный, ни-откуда see (Grot 1894, 60).

Orthography

The next necessary step in the normalization of a historical text is its orthographical normalization. That means that a uniform set of orthographic rules is applied to each text in order to render it homogenous and searchable. Below we will consider some of the issues connected with this task.

Establishing the set of automatic rules. By *automatic orthographic rule* we mean a rule which is applied to any string of symbols independently from its morphemic status. In modern Russian language the examples of such rules are writing the strings *чу, цу* with letter *y* instead of *ю*, and the strings *жи, ши* with letter *и* instead of *ы*.

If there are any alternations of graphemes in the same context, one of them must be chosen to be used in this context permanently. In the modern examples given above only the strings *чу, цу, жи* and *ши* are approved, while the homophonous strings *чю, цю, жы* and *шы* are considered to be clumsy errors. The same principle can be applied to strings in any texts, including those in LRC. Given such pairs as *'разсудѣиша' ~ 'разсудѣиша'*, *'бѣиша' ~ 'бѣиша'*, *'вѣдѣиша' ~ 'вѣдѣиша'*, *'приплѣиша' ~ 'приплѣиша'* etc., one can easily deduce that the differences within such pairs (where each member is the 3rd person plural active aorist form of the verbs *разсудити, быти, видѣти* and *приплѣити* correspondingly) are merely orthographic, and can be unified by applying the rule $[š'a] \rightarrow ша$ (or conversely $[š'a] \rightarrow шя$). Then such forms can be unified by means of a simple method like the following one (written in Python 3.6): `text = text.replace('шя','ша')`, where *text* is the name for the variable containing the text subject to normalization. A similar example is provided by the pair *щи ~ шы (сѹщимъ ~ сѹщымъ, преимѣиущихъ ~ преимѣиущихъ)*, and it can be dealt with after the same manner. Of course, there are a great deal of similar cases, and thus it is crucial to create a sufficient set of automatically applied rules, which could rule out the orthographical contradictions where it is possible to do so without human supervision.

Establishing the set of morpheme-based rules. Unfortunately, we have come to the point where the normalization issue overlaps with the annotation one. It is quite evident that a corpus without annotation is nearly useless for most specialists. A corpus must be at least POS-tagged and morphologically annotated. It is undoubtedly desirable that a corpus also contain a morpheme annotation, i. e., that words in the corpus are represented as lists of morphemes of which they consist. Then such lists could be transformed into a set of all morphemes found in the corpus. Many morphemes have more than one allomorph, and it is strongly desirable that the orthographic representation of the allomorphs have some basic principles. Thus, to represent the allomorphs correctly one has to establish the set of rules considering their morphological representation. Allomorphs can be defined either phonologically, or morphologically. For example, given a root $/(slad \sim slat) \sim slažd/$ we can say that the first two allomorphs are phonologically defined (*slad* comes before a voiced consonant or a vowel, *slat* comes before an unvoiced consonant), and the last is morphologically defined (Nida 1949, 44–45), because it appears only before a special subset of suffixes. We suppose that the phonologically defined allomorphs do not have to be orthographically distinguished at all, while the morphologically defined ones do. Thus, the two forms of the same root like in *рѣд-ко* and *рѣт-костю* have to be unified, being its phonological allomorphs: *рѣд-ко* and *рѣд-костю*. We also consider it reasonable to apply the same unification principle to suffixal and prefixal elements. Hence, sets (1a), (2a) and (3a) given below have to get the same orthographical representation (1b), (2b), (3b) of their phonologically defined allomorphs of the suffix *-id-* 'belonging to a certain geographical region' (Table 1).

to choose the maximal unit of division based on syntactic criteria, not an arbitrary string of words beginning with a capital letter and ending with a full stop. For this purpose, we propose a special unit called *block*. In terms of phrase structure, *block* is the phrase which is not dominated by (or, in other words, not included in) any other phrase. In terms of dependency structure, *block* is the dependency tree whose head is not dependent from any other head. In general, *block* is the tree which is not a subtree of any other tree.

Let us consider the beginning of the Latin text of “The Letter on the Moluccas” (Table 2).

Table 2. Some examples of the so-called blocks

LATIN SOURCE TEXT	RUSSIAN TRANSLATION
1. De Moluccis insulis, itemque aliis pluribus mirandis, quae nouissima Castellanorum nauigatio, Serenissimi Imperatoris Caroli V auspicio suscepta, nuper inuenit, Maximiliani Transyluani ad Reuerendissimum Cardinalem Saltzburgensem epistola lectu perquam iucunda.	1. О Молукидскихъ островѣхъ и ѳныхъ многѳхъ дѳвныхъ, иже новѳишее плаваніе кастеллановъ, рѣкше испанскихъ, потщаниемъ кротчайшаго самодержьца Карола пятаго събрано, еже ново обрѣте, Маѳимилиана Транъсилвана къ честнѳишему кардыналю салтызвурьенскому епистолиа краснѳиша чтѣніемъ.
2. REVERENDISSIME ac Illustrissime Domine, domine mi unice, humillime commendo.	2. Честнѳишии мнѣ и пресвѣтлѳишии владыко, владыко мой въжелѳннѳишии, <...>
3. Rediit his diebus una ex quinque illis nauibus, quas Caesar superioribus annis, dum Caesareae Augustae esset, in alienum et tot iam saeculis incognitum orbem miserat ad inquirendum insulas, in quibus aromata proueniunt.	3. Възвратѳлся ѣсть въ днѣхъ сіѳхъ единъ отъ пятихъ корабль онѳхъ, ѳже кесарь въ прежнихъ лѣтѣхъ, въ нихже кесарское управляше начѳлство, послѳлъ естъ въ страннии и толікими уже вѣки незнаемыи мѳръ къ разсмотрѳнію острововъ, въ нихже ражаются арамѳти.

Here the numbers 1, 2 and 3 mark the corresponding maximal syntactic units of the texts. Number 1 is the title. Number 2 is the address. Number 3 is a compound sentence. All three belong to different phrase categories, but all three are *blocks*, because for each of them there is no higher phrase which includes them (or no higher head from which their heads depend, in terms of dependency grammar).

Morphological processing and annotation

Inflectional morphology

The grammar annotation for a corpus must include lemmatization and grammar analysis. For a language featuring relatively high degree of variability in morphology, the inflectional model cannot be determined by an apriori set of rules taken from a kind of textbook: *it must be drawn from the corpus data*. Regarding the mixed Russian — Church Slavonic character of the language of Middle Russian translations from Latin, this is the only possible solution for the LRC texts, because it is not likely that any existing grammar description of Middle Russian or Church Slavonic could adequately account for the degree of variation found in real texts.

To extract the grammar data from a set of texts, one must assume a particular descriptive model. We tend to assume two closely related models for inflectional morphology description, namely that proposed by A. E. Polyakov for the Church Slavonic subcorpus of the Russian National Corpus (Polyakov 2014, 251) and that created by T. A. Arkhangelskiy and known as UniParser. Polyakov's model consists of two basic components:

- (1) a grammar dictionary;
- (2) a table of inflectional types (paradigms).

The grammar dictionary is a list of lexemes with information on their inflectional features. Each lexeme in the dictionary must contain the following information (Polyakov 2014, 251):

- lemma and its variants;
- POS tags;
- inflectional type or paradigm (represented by a certain paradigm code), and irregular inflectional forms.

Polakov claims that a dictionary entry may also contain some explanatory remarks on the meaning of rare words, but we regard it to be superfluous, mostly due to the fact that it doesn't have any connection with grammar. As mentioned before, Middle Russian paradigm types must be extracted by means of analyzing the set of texts, not by any existing language description. Hence there are some major steps constituting the process of creation of the inflectional morphology model for the Middle Russian part of LRC (Polakov 2014, 252–253):

1) First of all, it is necessary to create the list of inflectional forms found in the texts of the corpus. This is easily done, for example, by means of the following Python 3 program (where *text* is the variable for the set of texts in the corpus):

```
slovar = text.split(' ')
slovar_1 = sorted(set(slovar))
for i in slovar_1 :
    print(i)
```

Having a list of inflectional forms, we can sort it by desinence using the following program (where *text* is an alphabetic string of inflectional forms created by the previous program):

```
text_reversed = text[::-1]
text_reversed_2 = text_reversed.split(' ')
text_reversed_3 = sorted(set(text_reversed_2))
text_reversed_4 = str(text_reversed_3)
slovar = text_reversed_4[::-1]
slovar_2 = slovar.split(' ')
for i in slovar_2 :
    print(i)
```

2) Secondly, the most frequent words must be manually POS-tagged and lemmatized. We consider it reasonable to adopt the list of POS tags used in RNC Middle Russian corpus (Berdičevskis, Eckhoff, Gavrilova 2016, section 3.3.1.), which is with some slight modifications given in the POS tags chart (Table 3).

3) The next step is forming the inflectional classes for the lemmatized words. Though these classes have to be drawn from the corpus analysis, it is possible to borrow some basic principles of their representation (as well as some coinciding paradigms) from Polyakov's Church Slavonic grammar dictionary.

4) Finally, the annotator must apply the created paradigms to the rest of the inflectional forms, checking the results of the tentative automatic or semi-automatic analysis and improving them by manually correcting the wrong guesses of the analyzer.

Another way of producing a formalized description of a language accounting for the facts extracted from the corpus is the UniParser formalized grammar description format developed by T. A. Arkhangelskiy (Arkhangelskiy 2012), which is freely available here: <http://languedoc.philol.msu.ru:8082/fieldling/uniparser/>. UniParser allows for describing the grammar of a certain natural language as a set of UTF-8 plain text files. Its universal character (it is not designed for any particular language) lets us assume that it could be applied to our material as well.

As to the Latin part of LRC, it is unlikely that its morphology could feature any major deviations from the standard variant of this language, and that is why we consider it possible to use the existing information on the Latin inflectional types instead of thoroughly analyzing the inflectional morphology of the Latin texts in the corpus. The Uni-Parser format, being a universal tool for formalized language description, may be applied to Latin material too. In addition, there is also the so-called Classical Language Toolkit (CLTK) for Python 3.6 whose aim is to provide users with automatic processing tools for Greek and Latin. It may be used to lemmatize the Latin part of LRC as well.

Derivational morphology

At the moment we do not plan to annotate the derivational structure of lexemes in the corpus, but it may be a prospective task, especially regarding the fact that some of the words in Russian translations from Latin may feature one-to-one morphemic correspondences to their Latin prototypes, as in the following examples given by E. S. Fedorova (Fedorova 1999b, 90): *ab-surdum* → *о-глушено*, *ob-iectio* → *о-пирание* / *вз-споръ*, *con-venit* → *съ-идется* etc.

Syntax

Syntactic model

Before developing the fine properties of syntactic annotation for a corpus, one must choose a certain syntactic model whose principles would underlie the annotation structure. In general, there are two most widespread models of syntactic representation for natural language, namely dependency and constituency (Osborne 2014, 604). Both PROIEL and TOROT projects use a variant of dependency grammar (Haug et al. 2009, 27; Eckhoff, Berdičevskis 2016, 63). The developers argue that their choice was predominantly determined by the fact that the languages in both corpora had a rather free word order, so it would be convenient to use the formalism where word order information were kept out of the syntactic model (Haug et al. 2009, 27). We are not inclined to regard this property of dependency grammar (hereinafter DG) as an advantage; moreover, it seems to us that DG has a number of other disadvantages, especially

Table 3. POS tags used in RNC Middle Russian corpus

POS tag	Its meaning
A	Adjective
A-PRO	Adjective pronoun
ADV	Adverb
ADV-PRO	Pronominal/interrogative adverb
CONJ	Conjunction
INTJ	Interjection
N	Noun
N-PRO	Nominal pronoun
Q	Quantifier word/cardinal numeral
P	Preposition
V	Verb
D	Determiner

in comparison with some variants of phrase structure grammar (hereinafter PSG), based on constituency principle. Thus, we will continue this section as a gradual comparison of DG in the variant adopted for PROIEL and TOROT, and PSG in the variant of X-bar theory (Chomsky 2015, 45 ff.; Carnie 2008, 112–132).

Basic notions of the X-bar theory

There are some concepts which distinguish the X-bar theory from the basic variants of PSG. The first one is the notion of a head, which X-bar theory shares with dependency grammars. Each phrase is regarded as having a single head, i. e., the element which determines the syntactic properties of the whole phrase (Chomsky 2015, 47; Melchuk 2014, 13; Zwicky 1985). The head projects higher layers of structure, adding one element (each containing one dependent phrase) at a time (Carnie 2008, 120). The head of a phrase is an item of the lexicon; if a head item has substantive content, it is called *lexical* head; a head item without substantive content is called *functional* head (Chomsky 2015, 47–48). The following table (table 4) exhibits the most common categories of *lexical* heads and the corresponding phrases headed by these categories:

Table 4. The most common lexical heads

CATEGORY		PROJECTED PHRASE	
NAME	ABBREVIATION	NAME	ABBREVIATION
Noun	N	Noun phrase	NP
Verb	V	Verb phrase	VP
Adjective	A	Adjective phrase	AP
Adverb	Adv	Adverbial phrase	AdvP
Preposition	P	Prepositional phrase	PP
Complementizer	C	Complementizer phrase	CP

The phrases projected by different heads are believed to have the same inner structure. Let us substitute X, Y, Z or W for any head category. Then the inner structure of XP, i. e. the phrase headed by X, and the Phrase Structure Rules for its formation will be as in the following figure (fig. 5).

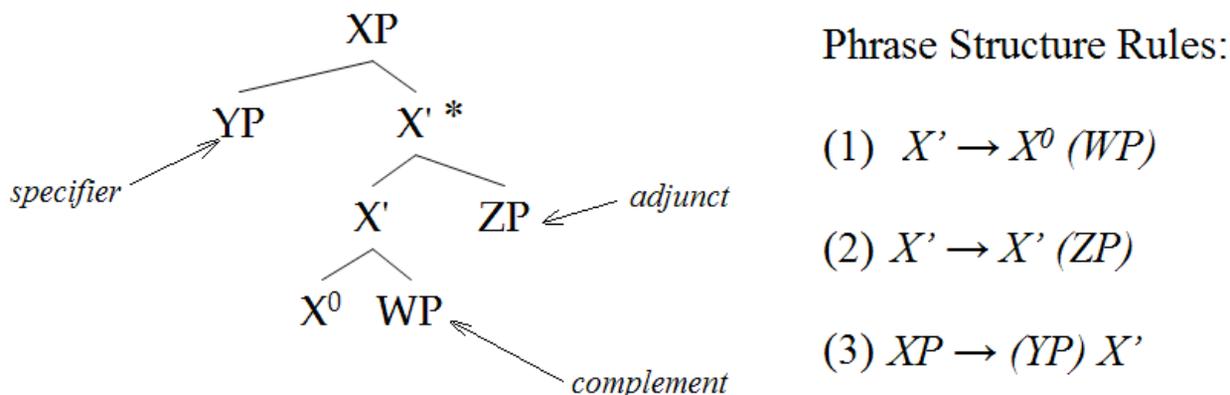


Fig. 5. The structure of XP and the corresponding phrase structure rules

The XP consists of several levels (Chomsky 2015, 48–49). The first level is formed by the head marked as X^0 and its sister phrase marked as WP, see rule (1). They form the first X' , i. e. the first X-bar level. The first X' level can add a ZP and form a new X' level, see rule (2). This rule is recursive; every X' level (called intermediate projection) is able to add a dependent phrase, thus forming a new X' level. The iterative character of the X' level is marked by an asterisk in the tree given above. Finally, there must be such an element YP, the addition of which finishes the construction of the phrase headed by X^0 , marking the whole construct as maximal projection XP, see rule (3). The layered character of a phrase allows for establishing the relations between its parts regarding their position. There are three basic relations, namely *complement-of*, *adjunct-of* and *specifier-of* (Chomsky 2015, 47–48), which we have marked by arrows in the above illustration. Below we give their definitions, taken from (Carnie 2008, 122).

A phrase that is a sister to a head is its *complement*. Phrases that are sisters to X' levels and are daughters of other X' levels are *adjuncts*. A phrase that is a sister to the X' level and a daughter of a maximal category (i. e. of XP) is a *specifier*. As we will see later, these relations can prove extremely useful in defining some crucial linguistic notions sometimes regarded as primitives.

Also note that phrase structure tree may as well be represented as a bracketed string of characters, see (fig. 6).

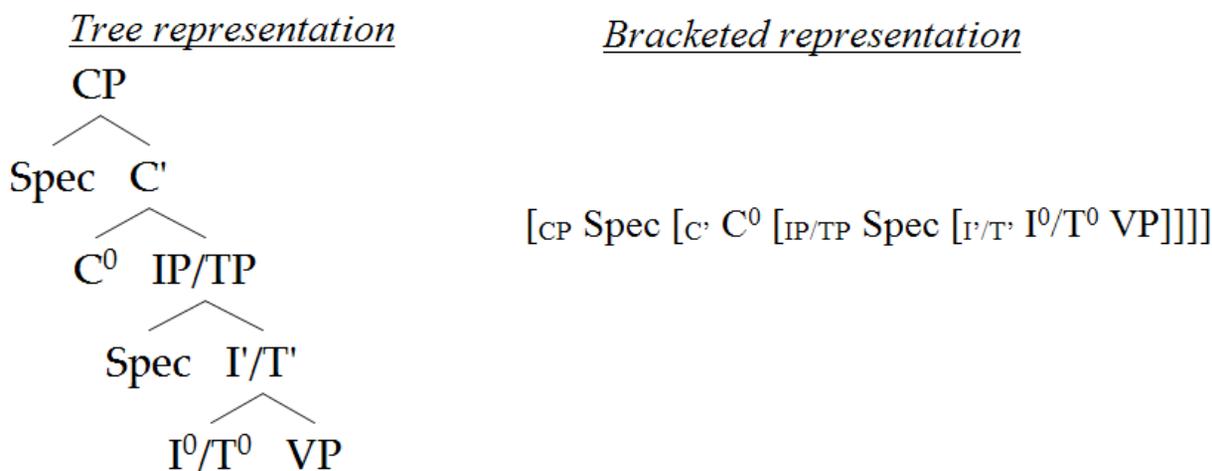


Fig. 6. Tree representation and bracketed representation

The other property of X-bar theory which is crucial for us is the idea of clause structure. Here we will give a brief outline of this structure, following (Chomsky 2015, 49). The full clause is assumed to be headed by a complementizer C hence being a CP. The complement of C is a propositional phrase headed by a functional category I (inflection) or T (tense), thus being IP or TP. I⁰ or T⁰ has the obligatory complement VP, which contains the predicate and its arguments:

The last crucial notion is the idea of *movement*, which means that the elements are able to change their initial position and move to a new one having left a *trace* (or a silent copy) of themselves in the position which they obtained initially. This simple mechanism has many advantages for the explanation of the linear order of elements and other relations in the tree. The displaced element and its traces are coindexed by a subscript letter (typically *i, j, k*) and linked by an arrow line. Of course, there is a number of other substantial notions which are characteristic of X-bar theory and other modules of generative grammar, but it would be superfluous to consider them in this proposal. Now, let us proceed to the comparison between the opportunities given by DG and the X-bar theory.

Subject, Object and Other Notions: A Comparison Between DG and PSG

The PROIEL DG annotation system allows for specifying each relation between two elements by ascribing it a particular type, e. g. *subject-of* (SUBJ), *object-of* (OBJ), *attribute-of* (ATR), *adverbial-of* (ADV) and so on, as demonstrated in an example from the Latin source of “The Letter on the Moluccas”, see (fig. 7)⁵:

Magellanus his dictis mirifice irritatus corrigit socios

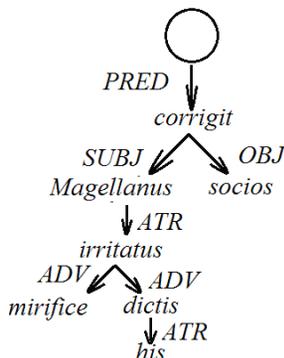


Fig. 7. Some syntactic relations in PROIEL

Subject, object and other relations are regarded here as linguistic primitives. In fact, they are not. Let us consider the following examples, where each subject and object is ascribed a semantic role (fig. 8).

All the examples demonstrate, that neither subjects nor objects can be associated with a single semantic role (also called theta-role). In addition, the examples (1) and (2) make it clear that such roles can be even opposite (*experiencer vs. stimulus, source vs. goal*). The example (3) shows that it is the construction, not the predicate itself, that defines the syntactic status of an argument: for example what would be an object in an active construction, becomes a subject in a passive one, as in (3b), preserving the same semantic role (*adducunt Serranum ~ Serranus adducitur*). Finally, the example (4) provides the evidence that subject cannot be associated with strictly one case, because it bears the nominative in a finite clause and the accusative in a non-finite one (which is traditionally called *accusativus cum infinitivo*).

⁵ All the examples of PROIEL annotation presented here are composed using the PROIEL guidelines. Due to some technical considerations they are not taken from actual PROIEL corpus texts.

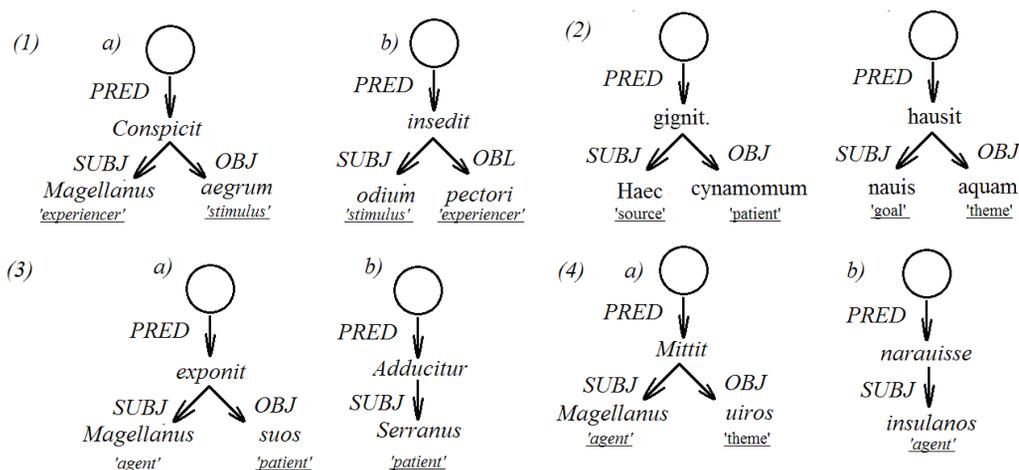


Fig. 8. The semantic roles of various subjects and objects

The DG system is not able to account for all these difficulties — as well as for some others, for example for the evident syntactic prominence of subject, see (McCloskey 1998, 197–198) — and its notions of subject, object and other syntactic relations stay vague and unclear. On the contrary, the X-bar system gives a precise and simple account of the given facts, which is based on the assumption that subject, object, attribute and so on are complex notions formed from a set of more simple relations. All we need for that are the notions of specifier, complement and adjunct, and the tripartite structure of a clause. The subject is taken to be a phrase base-generated in the specifier of VP, where the predicate ascribes it the semantic role (McCloskey 1998, 203–216). Then this phrase is raised into the position of the specifier of TP (frequently noted as [Spec, TP]) in order to get its case. If T^0 is finite, then the phrase in the specifier of TP gets nominative case; if T^0 is non-finite, it fails to ascribe the specifier of TP the nominative case, and it gets another case (accusative in Latin, dative in Slavic languages). So the subject is nothing but the specifier of TP (or IP). The object is, in its turn, the complement of VP (Chomsky 2015, 49). If the object phrase is raised to the position of the specifier of TP, as it usually happens during passivisation, it becomes a subject and gets the case associated with the subject of a particular clause (finite or non-finite). To sum up, the structural prominence of subject and the properties of object are explained by their syntactic position. That is the first fact that urges us to prefer the X-bar representation, like the one given below, to any possible DG annotation (fig. 9).

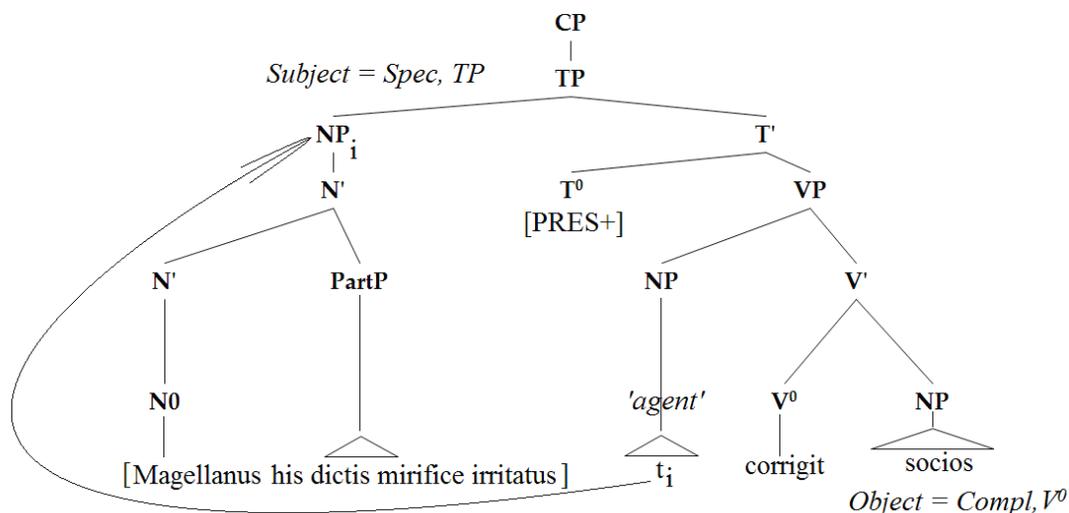


Fig. 9. Subject raising to the [Spec, TP] position in an X-bar tree

Relative Clauses and Movement

Now let us consider the following DG graph of the pair of syntactic units from the Latin-Russian parallel text of “The Letter on the Moluccas” (fig. 10).

regio, quam terram firmam uocant ~ *страна, юже и зѣмлю твёрду наричють*

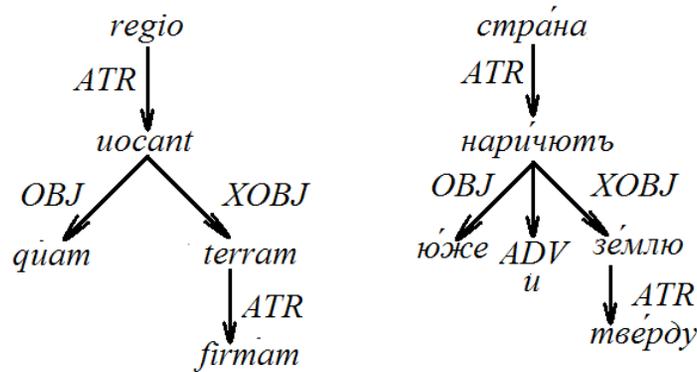


Fig. 10. A PROIEL-like DG representation of a relative clause

The annotation of this sample pair follows the guidelines given for relative clauses in PROIEL (Haug 2010, 38–43). One can see that the relative pronoun *quam* resp. *юже* is taken to be nothing but one of the dependents of the verb. It is really a dependent of the verb, but in addition it is the element which determines the properties of the whole relative clause: a relative clause lacking any relative element is a nonsense! But what determines the properties of a syntactic unit is the head of this unit. So, we come to a controversial situation: on the one hand, the relative element depends on the verb, being its argument; on the other hand, the same relative element is the head of the whole relative clause which contains the verb; to sum up, the relative pronoun holds two separate syntactic positions at the same time. As we can see, the PROIEL annotation cannot account for this problem. Some other types of DG annotation seem to be able to manage this problem a bit better, adopting some mechanisms of displacement, see (Osborne 2014, 619). Nevertheless, it seems to us that the simplest way is to adopt the well-known solution given by the generative grammar: namely, the idea that the relative pronoun is base-generated in a certain position related to predicate or the other element in the tree, and then moves to the leftmost position in the clause, becoming the specifier of a silent complementizer C:

regio, [*quam*_i C° [*terrām firmam uocant* t_i]] ~ *страна*, [*юже*_i C° [*и зѣмлю твёрду наричють* t_i]].

Word order representation and discourse configurationality

Latin, Church Slavonic and Middle Russian are *discourse-configurational languages*, i. e. the languages whose linear word order is predominantly determined by the information structure of utterances (Kiss 1995, 6). There seem to be at least two major notions associated with the informational structure of sentence, namely those of *topic* and *focus* (Kiss 1995, 6, 7–14, 15–24; Bailyn 2012, 266–267). The elements in discourse-configurational languages can be *topicalized* or *focalized*, that is moved to the positions associated with *topic* or *focus* (Bailyn 2012, 267).

We have already mentioned that DG graphs are usually independent from any particular linear order and thus are unable to represent the informational structure of utterances. On the contrary, the X-bar theory demands that the phrase marker for a particular sentence retain the linear order of its elements. That is why X-bar phrase markers allow for marking the informational structure of the elements by hosting them in special functional projections named Topic Phrase

(TopP) and Focus Phrase (FocP). Let us consider the following DG representation of a sentence taken from the Russian variant of “The Letter on the Moluccas” (fig. 11).

три́ сии́ о́строва вѣлие́ изобѣлие́ кариофи́лно но́сятъ

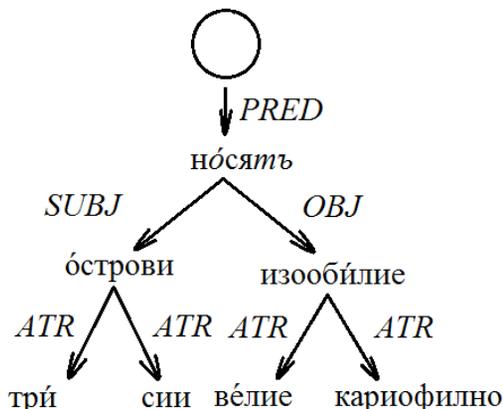


Fig. 11. A DG tree which conveys no information on linear order

The DG representation indicates the dependency relations between the elements; it also marks the type of each relation, but it still may correspond to a large number of possible linear orders:

- a. [Три сии острова] носятъ [изобилие велие кариофилно].
- b. Носятъ [три сии острова] [изобилие велие кариофилно].
- c. [Изобилие велие кариофилно] носятъ [три сии острова].
- d. [Изобилие велие кариофилно] [три сии острова] носятъ.
- e. [Изобилие кариофилно велие] [три сии острова] носятъ.
- f. [Три сии острова] [велие изобилие кариофилно] носятъ.

And so on, having at least $3! = 6$ positions for three elements {три, сии, острова} inside the subtree headed by the subject, the same number for those inside the subtree headed by the object and also $3! = 6$ positions for the subject, object and verb themselves. No part of the dependency tree can throw any light on the word order in the sentence.

It can be argued that some kinds of DG allow for including the word order information into the structure of tree-generating rules, like that described in (Hays 1964), but in fact such rules cannot cope with what is called *discontinuous* or *non-projective phrases*⁶, and, Middle Russian and Church Slavonic being discourse-configurational languages, with these languages themselves. Now let us consider the PSG representation of the same sentence in the form of X-bar phrase marker (fig. 12).

Here not only the linear order of elements is preserved as it is, but also the middle-field topicalisation of the noun phrase *велие изобилие кариофилно*, which results in SOV linear order (Bailyn 2012, 273–274), is marked as a syntactic operation (XP-movement). Hence one can easily decide that PSG has more explanatory force as to the linear and informational organization of utterances in discourse-configurational languages than DG.

⁶ *Discontinuous or non-projective phrases* are phrases which are linearly interrupted by some material which doesn't belong to the phrase.

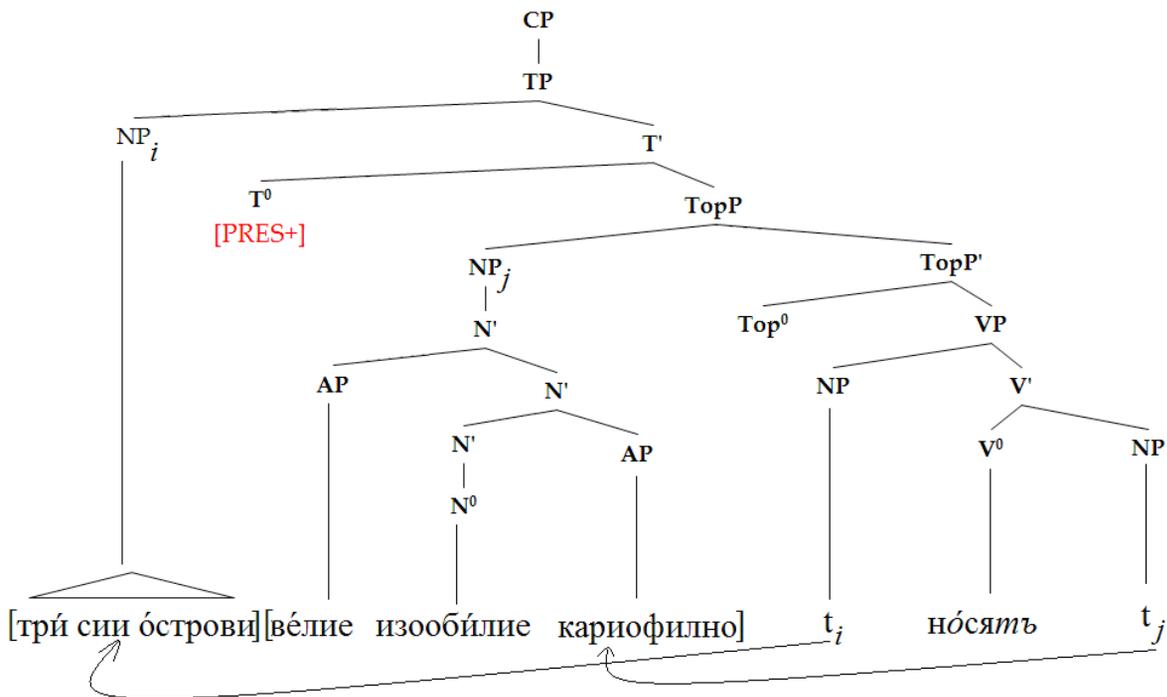


Fig. 12. A phrase marker of [Три сии острова] [велие избылие кариофилно] носятъ.

Which structure hosts more information?

To conclude with, now it is obvious that both DG and X-bar structures host the essential information about the dependency relations between the syntactic units, but the latter also has a number of mechanisms which can explain the linear order and information structure of utterances, as well as the complex nature of some basic linguistic concepts, while the former is unable to perform the same task. This simple reason convinces us that the X-bar theory is a better basis for corpus annotation than the DG system.

Syntactic annotation

At the current time we are not aware of any syntactic annotation tools for Latin or Middle Russian that use the PSG formalism. Latin and (Old, Middle, Modern) Russian corpora either lack syntactic annotation or are annotated in terms of dependency grammar. The only exception is the SKAT project, which includes some elements of phrase structure annotation in XML format, but its syntactic module is not ready for use so far (Alekseeva 2014), and in addition it is nonetheless based on the system of syntactic relations similar to that from the Russian National Corpus, i. e. dependency relations.

At any rate, there are some opportunities for establishing a PSG annotation system for these languages. These opportunities are based on two main facts:

(1) There exist a number of works exploring the limits and abilities of PSG description for old Indo-European languages, including Latin and old Slavic languages, for instance (Oniga 2014; Danckaert 2011) for Latin, (Mitrenina 2012) for Middle Russian, partly (Isakadze 1999) for Old Russian. Some of such works consider the problem of annotating the texts in these languages, e. g. see (Dimitrova 2011) for Old Church Slavonic.

(2) There are some open source syntax tree generators which let the users render a bracketed representation of a phrase into a syntax tree and back, for example:

Linguistic Tree Constructor (LTC): <http://ltc.sourceforge.net/about.html>

Syntax Tree Editor: <http://www.ductape.net/~eppie/tree/>

Besides, as we have already mentioned before, there exists a tool which allows for aligning the strings of word forms (Arkhangelskiy, Mishina, Pichkhadze 2014, 102). In addition, the Natural Language Toolkit for Python (NLTK) (Bird, Klein, Loper 2009, 291–326) supports some patterns of syntactic parsing which may be applied to our material too. All these instruments can be used as a base for our own syntax tree construction device included into LRC, but the question of the subtleties of their application remains quite unclear so far and requires a thorough analysis of the existing parsing techniques, tools and data. Of course, marking up the corpus after the generative manner requires quite a high level of expertise and is not a simple and straightforward task; that is why we hope to make final decision on the details of this procedure only after a tentative mark up of some text fragments.

Prospective directions of research

Grammatical and lexical variation in parallel texts

The relation between the sets of lexical items of the two languages cannot be reduced to a one-to-one correspondence. Particular lexical items are represented in the text only by the word forms, which are obligatory included in larger syntactic units called phrases. Thus, the comparison must begin not at the word level, but rather at the phrasal level considering the paired phrases from the source and target texts. The phrases in both parts of a parallel text must be divided into constituents, which must be paired using the appropriate software until the correspondences between the terminal nodes, i. e. the word forms, are established.

Let us consider three stages of analysis for an imaginary annotator illustrated below:

(1) Finding the phrase structure correspondences for the source and target phrase

In our example (see Fig. 12) the correspondences have to be established between two prepositional phrases: [*ad Taprobanen quam nunc Zamataram uocant*] and [*к Тапробáни, юже нынѣ Замáтару наричють*]. The corpus software must support manual syntactic annotation for phrases. The next step is the manual alignment of the corresponding nodes in the two phrase markers. Such alignment must look like that: $PP \rightarrow PP$, $P^o \rightarrow P^o$, $NP_i \rightarrow NP_i$, and so on, until the moment when each node of the Latin phrase marker is paired with the corresponding Russian node.

(2) Forming the list of lexical correspondences

After that the machine analyses the set of node pairs formed on the previous stage and offers to the annotator the corresponding set of paired word forms, like the one given on the second part of the illustration. If a word form corresponds to a terminal node, the annotator can relate it to a certain existing lemma or add a new one:

Taprobanen acc.sg. | Тапробан-е, N. f. 1gr. ~ Тапробáни acc.sg. | Тапробáни, Nf. indecl.

So for a pair of aligned terminal nodes of the source text and the target text one could derive a corresponding pair of lemmata:

Taproban-е, Nf. 1gr. → Тапробáни, Nf. indecl.

Such ordered pairs of lemmata $L(l,s) :=$ ‘the Latin lemma l corresponds to the Russian lemma s ’ may serve as the basis for an automatically formed vocabulary for each parallel text.

(3) Forming the strings of categorial symbols for a particular phrase

The strings of categorial symbols must be formed automatically, being simply the regular linearization of the Latin and Russian tagged phrase markers from the stage 1 (see fig. 13).

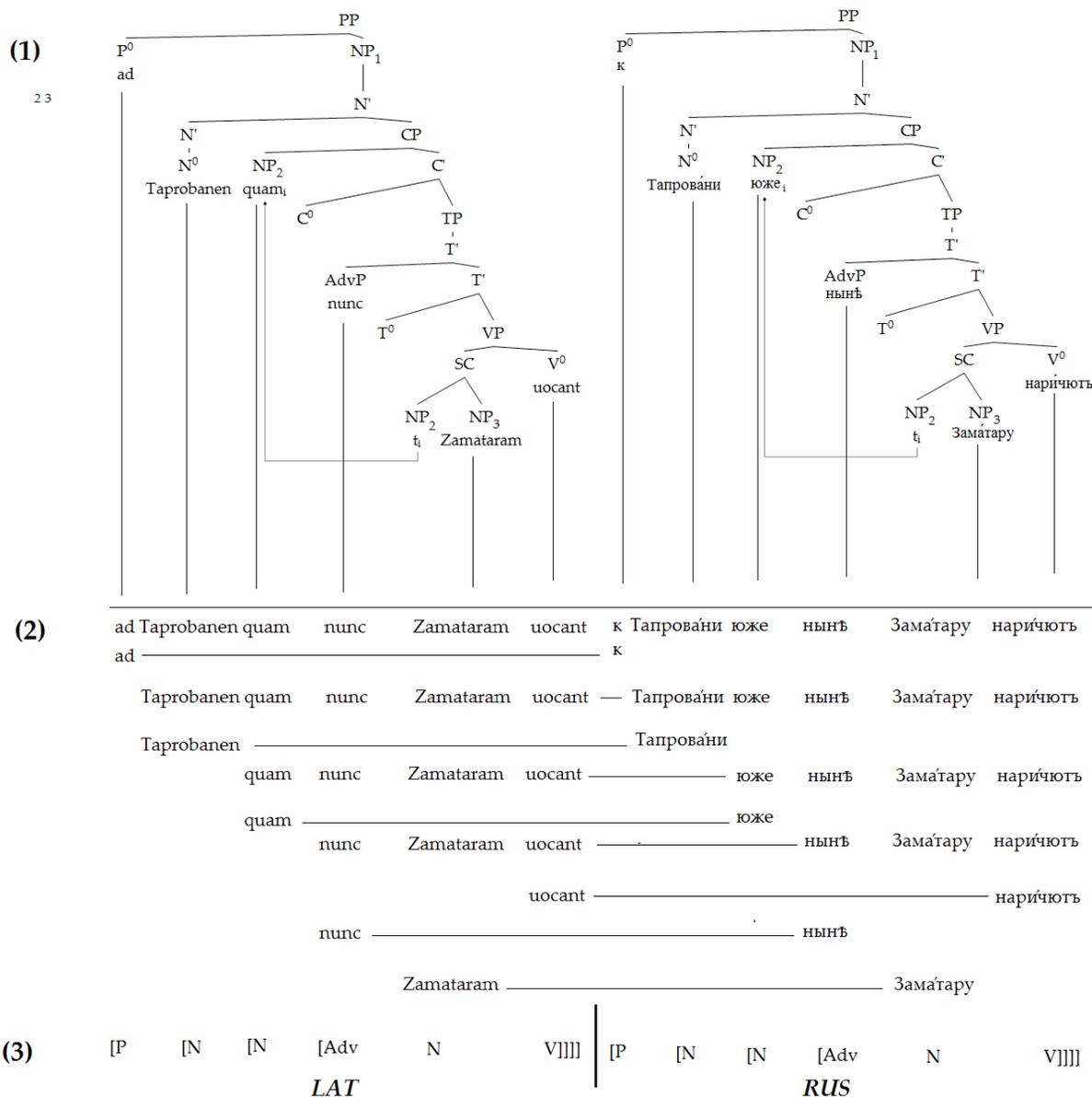


Fig. 13. An example of the lexical and phrasal correspondences in a particular bitext

Sometimes it is impossible to form a one-to-one correspondence between all terminal nodes of the source phrase and all terminal nodes of its target counterpart. For example, in a pair of correspondent noun phrases [*tropicum Capricornum*] ‘Tropic of Capricorn’ ~ [*вѣснаго солнечнаго възвращѣнiа*] ‘Tropic of Capricorn (a translator’s periphrase)’ it is possible to analyse the phrases themselves, but impossible to pair their constituents. In such cases the phrasal material must be stored in the vocabulary as a whole:

[_{NP} *tropicus Capricornus*] → [_{NP} *вѣсное солнечное възвращѣние*] In another case the Latin noun *Pigmeos acc.pl.* | *Pigmeus N m. 2* is translated by a substantivated adjective phrase [_{NP} ∅ [_{AP} *лакотныхъ возрастомъ*]]. The vocabulary entry for this pair must look like this:

Pigmeus N m. 2 → [_{NP} ∅ [_{AP} *лакотный возрастъ*]].

In other words, if on a certain stage there cannot be found any further one-to-one correspondences between the constituents of the Latin and Russian phrases, then the annotator has to enter into the vocabulary the last pair of correspondences, even if this pair contains nonterminal nodes.

Selective features of lexical items

The LCA project must also provide its users with an opportunity to research the lexical compatibility and selective features of the lexical items in the corpus.

Lexical compatibility search

Having an annotated and aligned phrase marker pair for a certain parallel text, one could easily find in it all the entries featuring a certain lexical head. One could also get a list of dependents it is compatible with, as well as the corresponding material in the opposite part of the same bitext.

Thus, the syntactically annotated corpus must give an opportunity to form a compatibility list for the lexical items in it or to supplement the existing vocabularies with such information.

Forming the subcategorization frames

A syntactically annotated corpus can be supplemented with a module serving for extraction of the subcategorization frames. If annotators tag the arguments of each predicate with their semantic role (theta-role), such a module will be able to compare this information with the syntactic position of the arguments and form a subcategorization frame for a certain predicate. Consider the pair of sentences in the (fig. 14).

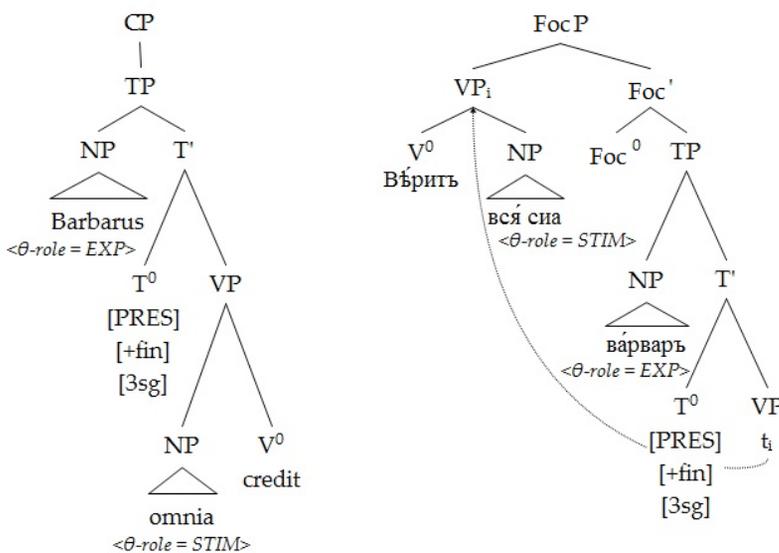


Fig. 14. A pair of sentences with the arguments of the predicate marked with their θ -roles

All the annotator has to do is to mark the subject in these sentences as experiencer (EXP) and the object as stimulus (STIM). Once the annotator does it, the module will form a certain subcategorization frame (Table 5).

Table 5. An example of corresponding subcategorization frames

LAT	RUS
Credo	вѣрити
EXP: Spec, CP (subject)	EXP: Spec, CP (subject)
STIM: Comp, V0 (object)	STIM: Comp, V0 (object)

Such subcategorization frame doesn't mark the structural cases of the arguments, because they depend on the clause type, not on the semantic role of a certain argument (for instance, as we have already mentioned, the subjects of finite clauses bear nominative case, while the subjects of infinite clauses bear accusative or dative).

Finding omissions and additions in parallel texts

It is typical of a Latin-Russian parallel text to include a number of omissions and additions. Translators (or later scribes) could remove large fragments, like in the example from the (Table 6), where the corresponding fragments are highlighted in bold type:

Table 6. An example of omission in translation

LAT	RUS
Nam ut reliqua omittam , tradidit Herodotus alioqui clarissimus autor, cynamomum in auium nidis reperiri, in quos uolucres illud ex longissimis regionibus, et praesertim Phoenix (cuius nidum nescio quis unquam uiderit) detulissent.	И да́ждь прóчая оста́влю.

The syntactic annotation must show where a nonempty set of nodes in a Latin text is conveyed by an empty set of nodes in its Russian counterpart. The search in the corpus or in a particular bitext must find all the results of that kind, letting the researcher assess the reasons and regular patterns of the omissions in a translated text.

There can be found some opposite cases, in which translators added something to the translation. In the example from the (Table 7) the addition is highlighted in bold type:

Table 7. An example of addition in translation

LAT	RUS
atque adeo Nili fontes et Troglodytas inuenerunt	И са́мья ни́ловы исто́чники, и глаголе́мья трогло́дѣти, ре́кше подь землею въ пеще́рахъ живу́щаа, обрѣ́тоша

The syntactic annotation must show such superfluous nodes in the translation, to ascertain the absence of any correspondences in the Latin source. Such option could also help the future researchers.

Marking the translator's faults

During the annotation process it is impossible not to find some translational mistakes. This can be regarded as an additional advantage of manually annotated parallel corpora. The translator's fault must be marked with a special metatag, allowing for a grammar note, which could possibly result in lists of typical translational mistakes for various parallel texts.

Conclusion

In this article we have presented a preliminary project of a deeply tagged parallel corpus of Russian translations from Latin, including the information on its goals, purposes, applicability

and structure. We have proposed annotation models for various levels of language representation, paying special attention to the issues of orthographical normalization, morphological and syntactic tagging of the corpus. The creation of such a corpus could provide researchers with a powerful instrument for scholarly activities in the fields of historical linguistics, literary studies and history of culture.

References

- Alekseeva, E. L. (2014) Sintaksicheskaya razmetka korpusa drevnerusskikh agiograficheskikh tekstov SKAT [Syntactic tagging of Saint-Petersburg corpus of hagiographic texts (SCAT)]. In: *Strukturnaya i prikladnaya lingvistika*. Iss. 10. Saint Petersburg: Saint Petersburg State University Publ., pp. 345–351. (In Russian)
- Arkhangelskiy, T. A. (2012) *Printsiipy postroeniya morfologicheskogo parsera dlya raznostrukturnykh yazykov. Extended abstract of PhD dissertation (Philology)*. Moscow, Moscow State University, 24 p. (In Russian)
- Arkhangelskiy, T. A., Mishina, E. A., Pichkhadze, A. A. (2014) Sistema elektronnoj grammaticheskoy razmetki drevnerusskikh i tserkovnoslavjanskikh tekstov i ee ispol'zovanie v veb-resursakh [A system for digital morphological tagging for Old Russian and Church Slavonic texts and its use in web resources]. In: V. A. Baranov, V. Zhelyazkova, A. M. Lavrent'ev (eds.). *Pismenoto nasledstvo i informacionnaya tehnologii. El'Manuscript–2014*. Sofia; Izhevsk: Bolgarskaya akademii nauk Publ., pp. 102–104. (In Russian)
- Bailyn, J. F. (2012) *The Syntax of Russian*. Cambridge; New York: Cambridge University Press, XVIII, 373 p. (In English)
- Berdičevskis, A., Eckhoff, H., Gavrilova, T. (2016) The beginning of a beautiful friendship: Rule-based and statistical analysis of Middle Russian. In: V. P. Selegej (ed.). *Computational linguistics and intellectual technologies: Proceedings of the International conference “Dialogue 2016”*. Vol. 15 (22). Moscow: Russian State University for the Humanities Publ., pp. 99–111. (In English)
- Bird, S., Klein, E., Loper, E. (2009) *Natural language processing with Python*. Beijing: O'Reilly, XX, 479 p. (In English)
- Carnie, A. (2008) *Constituent Structure*. Oxford; New York: Oxford University Press, XVIII, 292 p. (Oxford surveys in syntax and morphology. Book 5). (In English)
- Chomsky, N. (2015) *The Minimalist Program: 20. Anniversary edition*. Cambridge, MA: MIT Press, XIII, 393 p. (In English)
- Danckaert, L. (2011) *On the left periphery of the Latin embedded clauses. PhD dissertation (Philology)*. Ghent, Belgium, Ghent University, XVII, 387 p. (In English)
- Dimitrova, Ts. (2011) *The Old Bulgarian noun phrase: Towards an annotation specification*. Saarbrücken: VDM Verlag Dr. Müller, VII, 273, 28 p. (In English)
- Eckhoff, H. M., Berdičevskis, A. (2016) Automatic parsing as an efficient pre-annotation tool for historical texts. In: *Proceedings of the Workshop on language technology resources and tools for digital humanities (LT4DH)*. Stroudsburg, PA: The COLING 2016 organizing committee; Association for Computational Linguistics, pp. 62–70. (In English)
- Fedorova, E. S. (1999a) *Traktat Nikolaja de Liry “Probatio adventus Christi” i ego tserkovnoslavjanskij perevod kontsa XV veka*: In 2 books. Book 1. Moscow: Prosvetitel' Publ., 287 p. (In Russian)
- Fedorova, E. S. (1999b) *Traktat Nikolaja de Liry “Probatio adventus Christi” i ego tserkovnoslavjanskij perevod kontsa XV veka*: In 2 books. Book 2: *Prilozheniya*. Moscow: Prosvetitel' Publ., 120 p. (In Russian)
- Gaifman, H. (1965) Dependency systems and phrase-structure systems. *Information and Control*, 8 (3): 304–337. DOI: 10.1016/S0019-9958(65)90232-9 (In English)
- Gavrilova, T. S., Shalganova, T. A., Liashevskaja, O. N. (2016) K zadache avtomaticheskoy leksiko-grammaticheskoy razmetki starorususkogo korpusa XV–XVII vv. [Lexico-grammatical annotation of the Middle Russian corpus 1400–1700: A computational approach]. *Vestnik Pravoslav'nogo Svyato-Tikhonovskogo gumanitarnogo universiteta. Seriya III: Filologiya — St. Tikhon's University Review. Series III: Philology*, 2 (47): 7–25. DOI: 10.15382/sturIII201647.7-25 (In Russian)
- Grishman, R. (1999) Iterative alignment of syntactic structures for a bilingual corpus. In: S. Armstrong, K. Church, P. Isabelle et al. (eds.). *Natural language processing using very large corpora*. Dordrecht: Springer, pp. 225–234. (Text, Speech and Language Technology. Vol. 11.). DOI: 10.1007/978-94-017-2390-9_14 (In English)
- Grot, Ja. K. (1894) *Russkoe pravopisanie; Rukovodstvo, sostavlennoe po porucheniyu 2-go Otdeleniya Imperatorskoj akademii nauk akademikom Ya. K. Grotom*. 11th ed. Saint Petersburg: Tipografiya Imperatorskoj Akademii Nauk Publ., XII, 120, XL p. (In Russian)

- Haug, D. T. T. (2010) *PROIEL guidelines for annotation*. [Online]. Available at: https://folk.uio.no/daghaug/syntactic_guidelines.pdf (accessed 15.08.2019). (In English)
- Haug, D. T. T., Jøndal, M. L., Eckhoff, H. M. et al. (2009) Computational and linguistic issues in designing a syntactically annotated parallel corpus of Indo-European languages. *TAL (Traitement Automatique des Langues)*, 50 (2): 17–45. (In English)
- Hays, D. G. (1964) *Dependency theory: A formalism and some observations*. Santa Monica, CA: RAND Corporation, VII, 39 p. (In English)
- Isakadze, N. V. (1999) *Otrazhenie morfologii i referentsial'noj semantiki imennoj gruppy v formal'nom sintaksise. Extended abstract of PhD dissertation (Philology)*. Moscow, Moscow State University, 23 p. (In Russian)
- Kalugin, V. V. (2001) “Kniga svyatogo Avgustina” v russskoj pis'mennosti XVI — XIX vekov. In: A. M. Moldovan, V. S. Golysheko (ed.). *Lingvisticheskoe istochnikovedenie i istoriya russkogo yazyka*. Moscow: Drevlekhranilishche Publ., pp. 108–163. (In Russian)
- Kazakova, N. A. (1980) *Zapadnaya Evropa v russskoj pis'mennosti XV–XVI vekov. Iz istorii mezhdunarodnykh kul'turnykh svyazej Rossii*. Leningrad: Nauka Publ., 278 p. (In Russian)
- Kazakova, N. A., Katushkina, L. G. (1968) Russkij perevod XVI v. pervogo izvestiya o puteshestvii Magellana (Perevod pis'ma Maksimiliana Transil'vana). In: D. S. Likhachev (ed.). *Trudy otdela drevnerusskoj literatury*. Vol. 23. Leningrad: Nauka Publ., pp. 227–252. (In Russian)
- Kiss, K. É. (ed.). (1995) *Discourse Configurational languages*. New York; Oxford: Oxford University Press, 402 p. (Oxford Studies in Comparative Syntax). (In English)
- Kloss, B. M. (1975) Maksim Grek — perevodchik povesti Eneya Sil'viya “Vzyatie Konstantinopolya turkami” [Maxim the Greek — translator of Aeneas Silvius' narrative “Seizure of Constantinople by Turks”]. In: *Pamyatniki kul'tury. Novye otkrytiya. Pis'mennost', iskusstvo, arkheologiya*. Moscow: Nauka Publ., pp. 55–61. (In Russian)
- Matasova, T. A. (2014) Pervaya kniga “Geografii” Pomponiya Mely v drevnerusskom perevode: O retseptsii antichnogo naslediya v russskoj kul'ture XV–XVI vv. [The Old-Russian translation of the first part of Pomponius Melas' “Cosmography”: Perception of classical heritage in Russian culture in XV–XVI centuries]. *Aristej: vestnik klassicheskoy filologii i antichnoj istorii — Aristeeas. Philologia Classica et Historia Antiqua*, IX: 310–343. (In Russian)
- McCloskey, J. (1998) Subjecthood and subject positions. In: L. Haegeman (ed.). *Elements of grammar: Handbook in generative syntax*. Dordrecht: Springer, pp. 197–235. DOI: 10.1007/978-94-011-5420-8_5 (In English)
- Mitrenina, O. V. (2012) Sintaksis psevdokorrelyativnykh konstruksij s mestoimeniem *kotoryj* v starorussskom [The syntax of pseudo-correlative constructions with the pronoun *Kotoryj* (“Which”) in Middle Russian]. *Slověne. International Journal of Slavic Studies*, 1 (1): 61–73. DOI: 10.31168/2305-6754.2012.1.1.4 (In Russian)
- Mitrenina, O. V. (2014) The corpora of Old and Middle Russian texts as an advanced tool for exploring an extinguished language. *Scrinium. Journal of Patrology, Critical Hagiography, and Ecclesiastical History*, 10 (1): 455–461. DOI: 10.1163/18177565-90000109 (In English)
- Melchuk, I. (2014) Dependency in language. In: K. Gerdes, E. Hajičová, L. Wanner (eds.). *Dependency linguistics. Recent advances in linguistic theory using dependency structures*. Amsterdam; Philadelphia: John Benjamins Publishing Company, pp. 1–32. (Linguistik Aktuell / Linguistics Today. Vol. 215). (In English)
- Nida, E. A. (1949) *Morphology: The descriptive analysis of words*. Ann Arbor: University of Michigan Press, XVI, 342 p. (In English)
- Oniga, R. (2014) *Latin: A linguistic introduction*. Oxford: Oxford Universty Press, XVIII, 345 p. (In English)
- Osborne, T. (2014) Dependency grammar. In: A. Carnie, Y. Sato, D. Siddiqi (eds.). *The Routledge handbook of syntax*. Abingdon: Routledge, pp. 604–626. (In English)
- Partee, B. H., ter Meulen, A., Wall, R. E. (1990) *Mathematical methods in linguistics*. Dordrecht; Boston; London: Kluwer Academic Publishers, XX, 663 p. (In English)
- Polyakov, A. E. (2014) Korpus tserkovnoslavjanskikh tekstov: Problemy orfografii i grammatiki [Church Slavonic corpus: Spelling and grammar problems]. In: A. Kiklewicz (ed.). *Przegląd Wschodnioeuropejski [East European Review]*. Vol. V (1). Olsztyn: University of Warmia and Mazury in Olsztyn, pp. 245–254. (In Russian)
- Durandus, W. (2012) “*Rationale Divinorum officiorum*” Wilgelmi Durandi v russskom perevode kontsa XV veka. Moscow; Saint Petersburg: Indrik Publ., 261 p. (In Russian)
- Sokolov, E. G. (2014) “De moluccis insulis” Maksimiliana Transil'vana v russskom perevode XVI v.: Zadachi i perspektivy lingvisticheskogo issledovaniya [“De Moluccis Insulis” by Maximilianus Transylvanus in 16th century Russian translation: Tasks and prospects of the linguistic study]. *Vestnik Sankt-Peterburgskogo universiteta. Yazyk i literatura — Vestnik of Saint Petersburg University. Language and Literature*, 11 (3): 60–70. (In English)

- Tomelleri, V. S. (ed.). (1999) *Die "Pravila gramatichnye", der erste syntaktische Traktat in Rußland*. München: Verlag Otto Sagner, 159 p. (In German)
- Tomelleri, V. S. (2004) *Il Salterio commentato di Brunone di Würzburg in area slavo-orientale: Fra traduzione e tradizione (con un'appendice di testi)*. München: Verlag Otto Sagner, XVII, 343 p. (Slavistische Beiträge. Bd. 430). (In Italian)
- Tomelleri, V. S. (2011) *Latinskaya traditsiya u vostochnykh slavyan (nekotorye zametki)*. In: *Aktual'nye problemy filologii: Antichnaya kul'tura i slavyanskij mir*. Minsk: National Institute For Higher Education Publ., pp. 214–221. (In Russian)
- Tvorogov, O. V. (ed.). (1972) *Troyanskije skazaniya. Srednevekovye rytsarskie romany o Troyanskoj vojne po russkim rukopisyam XVI–XVII vekov*. Leningrad: Nauka Publ., 232 p. (In Russian)
- Tsyppin, D. O. (1990) *Skazaniye "O Molukitskykh ostrovekh" i Povest' o Loretskoj Bogomateri (Iz sbornika BAN, Arhangel'skoe sobr., D. 193, XVI v.)*. In: D. S. Likhachev (ed.). *Trudy otdela drevnerusskoj literatury*. Vol. 44. Moscow: Nauka Publ., pp. 378–386. (In Russian)
- Wimmer, E. (1990) *Die russisch-kirchenslavische Version von Maximilian Transylvans De Moluccis insulis ... epistola und ihr Autor. Zeitschrift für slavische Philologie*, 50 (1): 51–66. (In German)
- Wimmer, E. (2005) *Novgorod — ein Tor zum Westen? Die Übersetzungstätigkeit am Hofe des Novgoroder Erzbischofs Gennadij in ihrem historischen Kontext (um 1500)*. Hamburg: Kovac, 229 S. (Hamburger Beiträge zur Geschichte des östlichen Europa. Bd. 13). (In German)
- Zwicky, A. M. (1985) *Heads*. *Journal of Linguistics*, 2 (1): 1–29. DOI: 10.1017/S0022226700010008 (In English)
-

Author:

Evgenii G. Sokolov, ORCID: [0000-0001-5782-8093](https://orcid.org/0000-0001-5782-8093), e-mail: pan_liwerij@mail.ru

For citation: Sokolov, E. G. (2019) The project of a deeply tagged parallel corpus of Middle Russian translations from Latin. *Journal of Applied Linguistics and Lexicography*, 1 (2): 337–364. DOI: [10.33910/2687-0215-2019-1-2-337-364](https://doi.org/10.33910/2687-0215-2019-1-2-337-364)

Received 24 August 2019; reviewed 11 September 2019; accepted 12 September 2019.

Copyright: © The Author (2019). Published by Herzen State Pedagogical University of Russia. Open access under CC BY-NC License 4.0.

THE IMPACT OF SOME LINGUISTIC FEATURES ON THE QUALITY OF NEURAL MACHINE TRANSLATION

E. A. Shukshina✉¹

¹ Saint Petersburg State University, 7/9 Universitetskaya Emb., Saint Petersburg 199034, Russia

Abstract. This paper investigates how different features influence the translation quality of a Russian-English neural machine translation system. All the trained translation models are based on the OpenNMT-py system and share the state-of-the-art Transformer architecture. The majority of the models use the Yandex English-Russian parallel corpus as training data. The BLEU score on the test data of the WMT18 news translation task is used as the main measure of performance. In total, five different features are tested: tokenization, lowercase, the use of BPE (byte-pair encoding), the source of BPE, and the training corpus. The study shows that the use of tokenization and BPE seems to give considerable advantage while lowercase impacts the result insignificantly. As to the BPE vocabulary source, the use of bigger monolingual corpora such as News Crawl as opposed to the training corpus may provide a greater advantage. The thematic correspondence of the training and test data proved to be crucial. Quite high scores of the models so far may be attributed to the fact that both the Yandex parallel corpus and the WMT18 test set consist largely of news texts. At the same time, the models trained on the Open Subtitles parallel corpus show a substantially lower score on the WMT18 test set, and one comparable to the other models on a subset of Open Subtitles corpus not used in training. The expert evaluation of the two highest-scoring models showed that neither excels current Google Translate. The paper also provides an error classification, the most common errors being the wrong translation of proper names and polysemantic words.

Keywords: machine translation, neural machine translation, neural networks, transformer, translation evaluation, translation quality, tokenization, training corpus, byte-pair encoding, Yandex parallel corpus, Yandex corpus, WMT18 test set, news texts, Yandex.Translate, BLEU score.

Introduction

In 2016 Google launched its machine translation system based on neural networks (Turovsky 2016) that significantly improved the quality of translation. It was a milestone in the development of the field as shortly afterwards most translation companies were seeking to introduce it in their systems too. Next year, in 2017, Yandex.Translate also implemented neural networks.

A neural network consists of simple processors that can receive data, perform simple operations, and convey the result to other neurons. They are usually organized in layers: the input layer, the output layer, and the hidden layers in between them. The data is transmitted from one layer to the next in feed-forward neural networks, while recurrent neural networks (RNN) have 'loops' that enable information to be transmitted backwards as well.

Until the Transformer architecture was introduced in (Vaswani et al. 2017), the dominant neural machine translation models were based on RNN used in the encoder-decoder architecture (Sutskever, Vinyals, Le 2014) with an attention mechanism (Bahdanau, Cho, Bengio 2015) that vaguely corresponds to alignment.

The Transformer is based solely on the attention mechanism and does not employ a recurrent network structure. Just as earlier models, it consists of the encoding and decoding components. The novelty is in the use of 'self-attention' layers that allow to find connections between the words

in a sentence, and the ‘encoder-decoder attention’ layers that are concerned with the correspondence between the input and the output sequence.

The paper explores the five features the quality of translation depends on: the use of lowercase, tokenization, and BPE (byte-pair encoding), the source of BPE, and the training corpus.

Setup of the experiment

The models under study are based on the OpenNMT-py open machine translation system that provides a variety of tools for preprocessing the data as well as for training and testing translation models. The experiment is run on the Yandex en-ru bilingual corpus that contains one million aligned sentences automatically extracted from the web.

To evaluate the models, we use BLEU score on 3,000 test sentences of WMT18 news translation task (Bojar et al. 2018).

We compare several models that differ in the number of preprocessing steps that were applied to the training data:

1. Tokenization — provided by the Moses tokenizer distributed as a part of OpenNMT-py;
2. Lowercase;
3. BPE (Sennrich, Haddow, Birch 2016) — an approach to segment a text into subword units based on their co-occurrence frequencies.

All the models share the same transformer architecture with 6 layers of decoding and encoding inspired by (Vaswani et al. 2017) except for the multi-GPU feature that was not used in our setup.

Results

Table 1. BLEU scores for all possible combinations of three preprocessing steps: tokenization, lowercase, BPE (learnt from the training data)

Model	Tokenization	Lowercase	BPE	BLEU score
1	0	0	0	14.86
2	1	0	0	19.71
3	0	1	0	15.50
4	1	1	0	20.74
5	0	0	1	21.57
6	1	0	1	23.32
7	0	1	1	21.81
8	1	1	1	24.82

As we can see, the least helpful step of the three in question is lowercase for it increases the BLEU score of the system only by 0.85 points on average while tokenization and BPE have a much greater impact on the score increasing it on average by 3.7 and 5.2 points respectively.

Using a different corpus for learning BPE

The models with BPE presented above train BPE on the training data. However, its size may be insufficient to provide relevant vocabulary for the test data. The obvious step is to extract BPE vocabulary from larger monolingual corpora. This is supposed to provide a more general BPE vocabulary for each language that would not be specific to the training data.

For this purpose, we chose News Crawl with 8,233,935 sentences for Russian and 26,861,180 sentences for English. For better results, we deleted all the sentences that have no Cyrillic characters from the Russian News Crawl corpus, which reduced its size to 7,879,149 sentences.

Table 2 shows how 30,000 new BPE vocabularies impacted the tokenized and lowercased data.

Table 2. Results of the experiments with the BPE source and the training corpus. BLEU* stands for BLEU score measured on a part of OpenSubtitles corpus not used in training

Model	Tokenization	Lowercase	BPE	BLEU WMT 18	BLEU*
9	yes	yes	News Crawl	25.18	
10	yes	yes	News Crawl	17.85	26.50
11	yes	yes	Open Subtitles	16.02	25.72

Training on a bigger corpus

The size of the Yandex corpus is both an advantage, as it increases the speed of our experiments, and a disadvantage. To test performance on a bigger corpus, we used the Open Subtitles corpus that contains 25,910,105 Russian-English sentence pairs. We trained two models that differ in the source of the BPE vocabulary applied to the training data — News Crawl corpus for model 10 and the training data itself for model 11.

The results presented in Table 2 indicate that the use of a separate corpus for the extraction of BPE vocabulary proves to be more advantageous. The lower results of the models on the WMT 18 test data may be due to the difference in subject and register of the training and test data. The Open Subtitles corpus contains sentences of a more colloquial style, while WMT 18 test data is in line with the task through its focus on news text translation. To illustrate this difference, we also provide the BLEU score obtained on the test portion of the Open Subtitles corpus that was not used in training (BLEU* column of Table 2), that happens to be comparable to that of previous models.

Human evaluation

We decided to have a closer look at the translations provided by the two highest ranked models (models 9 and 10). For our evaluation we took 100 sentences randomly sampled from the test data. The performance of our models was compared to that of Google Translate. The raters were asked to rank the translations provided by the two models and Google Translate. They were given the input text and the reference translation from the test data. The rater could give the same rank to the sentences if they were equally good or bad. The results are presented in Table 3. It is clearly seen that neither of our models could excel Google translate.

Table 3. Results of human evaluation of models 9 and 10. Average comparison scores with Google Translate pairwise

	v9 vs Google	v10 vs Google	v9 vs v10
better	21	20	45
worse	49	56	30
equal	30	24	25

Error analysis

We also decided to examine the errors that occur in the translations of model 9 more thoroughly and sort them into the types that loosely correspond to the classification provided in (Vilar et al. 2006).

In the output of our model we found the following error types:

1. Missing words

a. Missing part of sentence: 12 sentences (8% of the total error count).

<i>Полчища владельцев прогулочных корабликов и артистов, изображающих Статую Свободы, линчевали бы его, если бы он попробовал это сделать.</i>	<i>A horde of boat-trip owners and Liberty impersonators would have lynched him if he did.</i>	<i>he would have done it if he tried to do it.</i>
--	--	--

b. Missing content words: 23 sentences (15% of the total error count).

<i>На мой взгляд, Коулмен <u>сегодня</u> один из самых выдающихся барабанищиков мира.</i>	<i>In my opinion, Coleman is one of the most accomplished drummers in the world <u>today</u>.</i>	<i>in my opinion, coleman is one of the world's most prominent drummers.</i>
---	---	--

c. Missing filler words: 13 sentences (8% of the total error count).

<i>Все в порядке, — шепчет одна <u>из</u> женщин.</i>	<i>It's okay, one <u>of</u> the women whispers.</i>	<i>everything is fine, one woman whispers.</i>
---	---	--

2. Incorrect words

a. Mistranslated proper names: 30 sentences (20% of the total error count). The majority were transliterated quite closely (*Ameliya Chesca* instead of *Amelia Chasse*, *Ranan Raffferti* instead of *Ronan Rafferty*). Some names were translated as if they were common nouns (*the wolves* instead of *Volkov*, *orphan* instead of *Sirotin*), and very few were far from correct (*Gennady Chelsi* instead of *Rod Chapel*).

b. Wrong sense of the word: 49 sentences (32% of the total error count).

<i>Мы также серьезно относимся к своему уставу и к власти, который он нам дает».</i>	<i>We also take the statute and the authority it gives us seriously.”</i>	<i>we are also seriously <u>concerned with</u> our charter and the power it gives us.”</i>
--	---	--

c. Wrong form of the word: 13 sentences (8% of the total error count).

<i>...но он получил повестку, обязывающую его явиться в отделение полиции...</i>	<i>...but he received a summons obliging him to appear at the police station...</i>	<i>...but he was given <u>a</u> agenda that would oblige him to appear in the police department...</i>
--	---	--

3. Word order errors — including word and phrase level reordering — were found in 13 sentences and correspond to 8% of the total error count.

<i>По словам Пола, и Люк, и Марк были «недовольны финансовыми условиями своего отделения».</i>	<i>Both Luke and Mark had become, Paul says, “bitter about the terms of their financial separation.”</i>	<i><u>paul and luke said that mark</u> were “dissatisfied with the financial conditions of his division.”</i>
--	--	---

Conclusion and future work

We investigated the impact of five different features on the quality of neural machine translation. The application of tokenization and BPE leads to a drastic growth in BLEU score. It is more effective to use larger monolingual corpora for BPE training. The use of lowercase does not seem to provide the advantage significant enough to compensate for missing capitalization

in proper names, abbreviations and beginnings of sentences. The study shows that thematic and register correspondence between the training corpus and the intended use of the system is quite important. This implies that a general-purpose translation system must be trained on a large representative parallel corpus with texts in different styles and registers as well as a wide range of topics.

It is worth mentioning that these conclusions are drawn from a single study based on Russian-English translation. All the statements remain to be verified for other language pairs, which is something we will focus on in our future work.

Sources

- Anglo-russkij parallel'nyj korpus (versiya 1.3)*. [Online]. Available at: <https://translate.yandex.ru/corpus> (accessed 15.08.2019). (In Russian)
- Index of /news-crawl*. [Online]. Available at: <http://data.statmt.org/news-crawl/> (accessed 11.09.2019). (In English)
- OpenSubtitles.org*. [Online]. Available at: <http://www.opensubtitles.org> (accessed 13.08.2019). (In Russian)

References

- Bahdanau, D., Cho, K., Bengio, Y. (2015) Neural machine translation by jointly learning to align and translate. *arXiv:1409.0473v7*. [Online]. Available at: <https://arxiv.org/abs/1409.0473> (accessed 15.08.2019). (In English)
- Barrault, L., Bojar, O., Costa-jussà, M. R. et al. (2019) Findings of the 2019 Conference on Machine Translation (WMT19). In: *Proceedings of the Fourth Conference on Machine Translation (WMT). Vol. 2: Shared Task Papers (Day 1). Florence, Italy, August 1–2, 2019*. Stroudsburg, PA: Association for Computational Linguistics, pp. 1–61. (In English)
- Bojar, O., Federmann, Ch., Fishel, M. et al. (2018) Findings of the 2018 Conference on Machine Translation (WMT18). In: *Proceedings of the Third Conference on Machine Translation (WMT). Vol. 2: Shared Task Papers. Brussels, Belgium, October 31 – November 1, 2018*. Stroudsburg, PA: Association for Computational Linguistics, pp. 272–307. (In English)
- Lison, P., Tiedemann, J. (2016) OpenSubtitles 2016: Extracting Large Parallel Corpora from Movie and TV Subtitles. In: *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC 2016). Portorož, Slovenia, May 23–28, 2016*. Pp. 923–929. [Online]. Available at: <http://www.lrec-conf.org/proceedings/lrec2016/summaries/947.html> (accessed 13.08.2019). (In English)
- One model is better than two. Yandex.Translate launches a hybrid machine translation system. (2017) *Yandex Blog*. 14 September. [Online]. Available at: <https://yandex.com/company/blog/one-model-is-better-than-two-yu-yandex-translate-launches-a-hybrid-machine-translation-system> (accessed 15.08.2019) (In English)
- Sennrich, R., Haddow, B., Birch, A. (2016) Neural Machine Translation of Rare Words with Subword Units. In: *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL 2016). Berlin, Germany, August 7–12, 2016*. Vol. 1. Stroudsburg, PA: Association for Computational Linguistics, pp. 1715–1725. (In English)
- Sutskever, I., Vinyals, O., Le, Q. V. (2014) Sequence to Sequence Learning with Neural Networks. In: *Advances in Neural Information Processing Systems 27 (NIPS 2014)*. Red Hook, NY: Curran Associates, pp. 3104–3112. (In English)
- Turovsky, B. (2016) Found in translation: More accurate, fluent sentences in Google Translate. *Translate. News about Google Translate*. 15 November. [Online]. Available at: <https://www.blog.google/products/translate/found-translation-more-accurate-fluent-sentences-google-translate/> (accessed 15.08.2019). (In English)
- Vaswani, A., Shazeer, N., Parmar, N. et al. (2017) Attention is all you need. In: *Advances in Neural Information Processing Systems 30 (NIPS 2017). Long Beach, California, USA, 4–9 December 2017*. Red Hook, NY: Curran Associates, pp. 5998–6008. (In English)

Vilar, D., Xu, J., D'Haro, L. F., Ney, H. (2006) Error analysis of statistical machine translation output. *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC-2006), Genoa, Italy, May 22–28, 2006*. Pp. 697–702. [Online]. Available at: http://www.lrec-conf.org/proceedings/lrec2006/pdf/413_pdf.pdf (accessed 10.08.2019). (In English)

Author:

Elena A. Shukshina, ORCID: [0000-0002-6014-9136](https://orcid.org/0000-0002-6014-9136), e-mail: elena.shukshina@gmail.com

For citation: Shukshina, E. A. (2019) The impact of some linguistic features on the quality of neural machine translation. *Journal of Applied Linguistics and Lexicography*, 1 (2): 365–370. DOI: 10.33910/2687-0215-2019-1-2-365-370

Received 29 August 2019; reviewed 11 September 2019; accepted 12 September 2019.

Copyright: © The Author (2019). Published by Herzen State Pedagogical University of Russia. Open access under CC BY-NC License 4.0.

УСТАРЕВШАЯ ЛЕКСИКА РУССКОГО ЯЗЫКА: ВОПРОСЫ ПРЕПОДАВАНИЯ И ЛЕКСИКОГРАФИЧЕСКОЙ ИНТЕРПРЕТАЦИИ

Е. В. Генералова ^{1,2}

¹ Российский государственный педагогический университет им. А. И. Герцена, 191186, Россия,
Санкт-Петербург, наб. реки Мойки, д. 48

² Санкт-Петербургский государственный университет, 199034, Россия,
Санкт-Петербург, Университетская наб., д. 7/9

OBSOLESCENT VOCABULARY OF THE RUSSIAN LANGUAGE: EDUCATIONAL AND LEXICOGRAPHIC INTERPRETATION ISSUES

E. V. Generalova ^{1,2}

¹ Herzen State Pedagogical University of Russia, 48 Moika River Emb., Saint Petersburg 191186, Russia

² Saint Petersburg State University, 7/9 Universitetskaya Emb., Saint Petersburg 199034, Russia

Аннотация. Статья посвящена рассмотрению устаревшей лексики в прикладном отношении (лексикографическом и учебном аспектах). Нерешенность теоретических вопросов осложняет практическое изучение, преподавание и словарное описание этого лексического массива; с другой стороны, на современном этапе развития науки теоретические лингвистические задачи во многом решаются именно средствами лексикографии. В учебном процессе важно не только семантизировать устаревшие лингвистические единицы, встречающиеся в классической литературе, но и дать представление учащимся о системности обновления словарного состава, о динамичности и историчности процесса выхода слов и устойчивых выражений из активного состава языка. Необходимо обратить внимание на разную степень устаревания лексики, на характерные для развития русского языка на современном этапе явления деактуализации и актуализации лексики, а также обращаться в учебном процессе к словарям: редких и устаревших слов, толковым, желателен историческим. Толковый словарь должен быть использован в обучении именно как источник изучения языка, т. к. лексикографическая интерпретация устаревшей лексики дает представление о таких словах и выражениях как части лексического состава русского языка. Основные проблемы толковой лексикографии в отношении представления слов и выражений, вышедших из активного состава языка: критерии включения устаревшей лексики в словарь и принципы ее описания (система помет и типовые

Abstract. The research is devoted to lexicographic and educational issues associated with obsolescent words in the Russian language. Unresolved theoretical issues, i. e. a classification of obsolete vocabulary, research on obsolete words in connection with the chronology of their use in the language, etc., complicates applied research, teaching and lexicographical description of such vocabulary. Within the educational process it is important not only to interpret the meaning of obsolete lexical units found in classical literature, but also to enable students to understand the systematic nature, the dynamism, and the historical character of the vocabulary evolution phenomenon. It is necessary to consider the different degree of obsolescence and the processes of vocabulary update typical for the development of the modern Russian language. The author suggests that dictionaries should also be used in the process, i. e., dictionaries of rare and obsolete words, explanatory dictionaries, and, additionally, historical dictionaries. The features of the lexicographic presentation of obsolete vocabulary should also be taken into consideration. The central issues of explanatory lexicography in this regard are the following: criteria for including obsolete lexis in dictionaries, and principles of its description (a system of lexicographic notes and typical explanations, various interpretations of historicisms and archaisms, etc.). The article discusses the lexicographic practice of academic dictionaries (in particular, the “Dictionary of the Modern Russian Language”) and the innovative project titled “The Dictionary of the Russian Language of the 21st Century”, supervised

толкования, вопрос различения и разной интерпретации историзмов и архаизмов). В статье рассматривается лексикографическая практика проектов, являющихся самыми новыми (т. е. наиболее близкими современному состоянию языка) и самыми подробными (толковыми словарями большого типа), — «Словаря современного русского языка» (БАС-3) и «Словаря русского языка XXI в.», создаваемого под руководством Г. Н. Складневской. Существенным достоинством современного антропоцентрически ориентированного толкового словаря является показ места и особенностей функционирования устаревшей лексики в современном языке, т. е. интерпретация этой лексики как неотъемлемой части лексического состава русского языка. Отвечающее современному уровню развития науки преподавание тем, связанных с устаревшей лексикой, и доступное пользователю описание этой лексики в современных словарях русского языка — важные предпосылки понимания и правильного использования в речи устаревших слов и выражений. При этом дискуссионные теоретические вопросы, связанные с исследованием устаревшей лексики, могут быть решены во многом в процессе ее словарного описания.

Ключевые слова: устаревшая лексика, словари редких и устаревших слов, толковые словари, исторические словари, лексикографические пометы.

Устаревшая лексика как существенная часть словарного состава языка всегда привлекала внимание исследователей, однако сам статус этого пласта лексики, его стратификация, представление в словарях активно обсуждаются и в настоящее время. На современном этапе развития лексикологии русского языка при значительном количестве разнообразных исследований, посвященных устаревшим словам и выражениям русского языка, остается нерешенным ряд теоретических и практических вопросов. Одним из основополагающих методологических аспектов является уточнение классификации устаревшей лексики. В настоящее время осуществляются попытки дополнить или детализировать традиционную классификацию выделения разрядов устаревшей лексики в зависимости от причин устаревания слов: так, помимо архаизмов различного типа и историзмов, отдельные исследователи выделяют фразеологические архаизмы (Попов 1995, 88–90), субстраты (вышедшие из употребления языковые элементы, закрепленные на периферии словарного состава, непродуктивные с точки зрения языковой ценности и не обладающие прозрачной внутренней формой) (Аркадьева и др. 2014, 25–26), нотиолизмы (не имеющие однословных соответствий в современном языке лингвистические единицы, вышедшие из употребления по причине утраты соответствующих понятий, при том что сами реалии сохранились) (Норман 2016), асимметричные архаизмы (устаревшие слова, замененные в современном русском языке не словами, а словосочетаниями) (Правдина, Чуриков 2016) и др. Исследователями также поставлен важный вопрос об изучении устаревшей лексики в зависимости от хронологических пластов ее существования в языке (должны ли, например, советизмы исследоваться по тем же принципам, что и устаревшая лексика, восходящая к XVI–XVII вв.?). На настоящем этапе развития русского языка злободневно исследование процессов деактуализации лексики и, напротив, актуализации, возвращения утраченных ранее активным фондом языка слов и фразеологизмов, а также оживления исторической

by G. N. Sklyarevskaya. The author suggests that an essential advantage of the explanatory dictionary is that it can define the role and features of outdated vocabulary functioning in the modern language, interpreting this vocabulary as an integral part of the lexical composition of the Russian language. The author concludes that studying obsolete vocabulary at school and making its descriptions available for users in modern Russian language dictionaries are important prerequisites for correct interpretation and use of obsolete words and expressions.

Keywords: outdated vocabulary, dictionaries of rare and obsolete words, explanatory dictionaries, historical dictionaries, lexicographic notes.

памяти слова, поскольку это активный процесс, проходящий именно в русской лексике XXI в.

Устаревшая лексика изучается и с точки зрения прикладной лингвистики: в отношении стилистических функций вышедших из активного запаса слов (в языке и в творчестве отдельных авторов), этнокультурного компонента, сложности перевода и преподавания (в частности, иностранцам); отдельным предметом изучения является словарное представление устаревшей лексики.

Настоящая статья посвящена рассмотрению устаревшей лексики в прикладном отношении, а конкретно — в лексикографическом и учебном аспектах.

Актуальность профессиональной и по возможности доступной широкому кругу носителей языка интерпретации устаревшей лексики не вызывает сомнения: правильное использование этих слов и выражений определяет уровень общей культуры и владения языком, и речь идет о значительном лексическом массиве, границы которого с трудом определяются лингвистами. Знание и употребление этой лексики в речи напрямую зависят от образования, возраста, социального статуса носителя языка; в частности, владение устаревшей лексикой учащимися далеко от совершенства и в отношении количества устаревших слов и выражений, входящих в активный и пассивный словарь студента или школьника, и в отношении использования этих слов в контекстах, исключающих лексическую ошибку. Во многом практическое изучение, преподавание и словарное описание устаревшей лексики осложнено нерешенностью теоретических вопросов. С другой стороны, именно лексикографическая интерпретация устаревшей лексики в современных словарях может способствовать прояснению дискуссионных моментов.

В школьном образовании изучение устаревшей лексики состоит в констатации существования разных временных пластов в составе современного русского языка и разных разрядов (архаизмов и историзмов) в составе устаревшей лексики и в лучшем случае подкрепляется семантизацией устаревших слов из произведений классической русской литературы.

Практика показывает, что речь, как правило, идет об устаревшей лексике XVIII–XIX вв., т. е. о словах и выражениях, встречающихся в произведениях школьной программы; в меньшей степени привлекаются так называемые давно устаревшие или «старинные слова» (термин был введен Н. М. Шанским (Шанский 1972, 145) для обозначения лексем, относящихся к отдаленным по времени эпохам) и, наоборот, слова и выражения, относительно недавно вышедшие из употребления, в связи с чем современный школьник часто не понимает, что к разряду устаревших слов относятся, например, и советизмы (*колхоз, пятилетка, стахановец*), и недавно перешедшие в пассивный запас слова (*пейджер, аська*). Представляется необходимым обращать особое внимание учащихся на то, что устаревание слова связано с тем, к какой тематической группе принадлежит лексическая единица, что устаревшие слова относятся к разным эпохам и понятия «устаревшее/новое» относительно, являются историческими, при этом процесс устаревания разных слов идет с разной скоростью. Важно дать представление о том, что выход слова из активного состава языка — динамический и исторический процесс, в результате чего в языке выделяются не только устаревшие, но и устаревающие слова и выражения (например, *баловень, барыш, мор, электронно-вычислительная машина*). Т. е. степень устаревания слова может быть разной — от полного исчезновения (например, без обращения к специальной справочной литературе нельзя узнать, что слово *варворка* означало ‘шарик на конце кисти’¹) до перехода в пас-

¹ Для таких слов с непрозрачной морфемной структурой Т. Г. Аркадьевой и коллегами (Аркадьева и др. 2014) и был предложен термин «субстракт». Введение самого термина может быть обсуждаемо, но нельзя не согласиться с исследователями, что эта категория слов, несомненно, заслуживает выделения и должна быть пред-

сивный словарный запас (*сказание, лампада, уста*). Особое внимание следует уделить явлениям деактуализации и актуализации лексики, поскольку это активные процессы, характерные для развития лексического состава русского языка на современном этапе.

Чтобы сделать эту тему интересной, помимо семантизации устаревших слов, поиска их в текстах упражнений и определения их типа (а классификация может быть расширена с учетом новейших лингвистических разработок), школьникам могут быть предложены задания разного вида: соотнести устаревшие слова и фразеологизмы и их значения (в том числе старые и новые топонимы), найти историзмы в тексте и сделать вывод о том, к какой тематической группе относятся эти слова (или с помощью исторических и толковых словарей распределить историзмы по тематическим группам), подписать на рисунках устаревшие названия цветов в радуге, названия частей тела человека, обозначения вооружения и т. п., найти лишнюю единицу в каждой группе и т. п. Отдельно стоит обратить внимание на судьбу устаревших слов в современном языке: они используются в пословицах и поговорках (*аршин проглотил, семи пядей во лбу*), в художественной литературе при описании событий истории, для обозначения реалий заграничной жизни (*престолонаследник, герцог*), в современной речи с иронической окраской (*школьное вече, велеречивый начальник*), деактуализируются (*гимназия, губернатор*).

В учебном процессе необходимо использование достижений лексикографии: привлечение словарей редких и устаревших слов, толковых словарей, желательны также исторических словарей. Основная функция специальных словарей редких и устаревших слов — помощь в чтении произведений русской классической литературы, т. е. устранение непонимания между людьми разных эпох, поэтому такие словари в первую очередь адресованы учащимся и студентам. На настоящий момент существует много интересных словарей редких и устаревших слов, которые, являясь лингвистическими и культурологическими справочниками, должны активно использоваться в процессе обучения: см. Р. П. Рогожникова, Т. С. Карская «Школьный словарь устаревших слов русского языка: по произведениям русских писателей XVIII–XX вв.» (Рогожникова, Карская 1996), Н. Г. Ткаченко, И. В. Андреева, Н. В. Баско «Словарь устаревших слов» (Ткаченко, Андреева, Баско 1997), В. С. Елистратов «Язык старой Москвы: лингвоэнциклопедический словарь» (Елистратов 1997), В. П. Сомов «Словарь редких и забытых слов» (Сомов 1996), Л. А. Глинкина «Иллюстрированный словарь забытых и трудных слов из произведений русской литературы XVIII–XX вв.» (Глинкина 1998), Ю. А. Федосюк «Что непонятно у классиков, или Энциклопедия русского быта XIX века» (Федосюк 2000), С. К. Бирюкова «Словарь культуроведческой лексики русской классической литературы» (Бирюков 2003) и др.

Обращение к историческим словарям в школьной программе проводится редко, хотя историческая лексикография русского языка находится на очень высоком научном уровне. Знакомство школьников с основными историческими словарями русского языка, представляющими собой многотомные продолжающиеся издания («Словарь русского языка XI–XVII вв.» (Сл11–17), «Словарь русского языка XVIII в.» (Сл18), «Словарь обиходного русского языка Московской Руси XVI–XVII вв.» (СОРЯ)), всегда вызывает интерес учащихся и расширяет их представления о разнообразии существующих справочных изданий и филологии в целом. В исторических словарях учащиеся смогут найти информацию о существовании определенных слов и их значениях в русском языке конкретного периода; на основе этих словарей школьники могут, например, выделять типы устаревших слов: историзмы и различные типы архаизмов (фонетические архаизмы: *воксал* (ср. *вокзал*), *осьмнадцать* (ср. *восемнадцать*), *номер* (ср. *номер*), словообразовательные архаизмы:

ставлена в преподавании и в словарях по другим принципам, чем встречающиеся в литературе историзмы и легко семантизируемые архаизмы.

рыбарь (ср. *рыбак*), *содейство* (ср. *содействие*), собственно лексические архаизмы: *одр* 'постель', *оний* 'тот', семантические архаизмы: *гость* 'иностранный купец', *поезд* 'конный отряд', *живот* 'жизнь'). Учащимся также может быть предложено задание написать рассказ с использованием устаревшей лексики, при этом желательно обратить внимание школьников на функции устаревшей лексики в языке: историзмы используются в исторических и художественных текстах для точного описания уже исчезнувших явлений, архаизмы встречаются в основном для придания исторической достоверности речи персонажей, а также для создания особого стилистического эффекта — торжественно-высокого или, напротив, иронического.

При изучении устаревшей лексики принципиально важно обращение к толковым словарям. В данном случае значимо, что современный толковый словарь не просто справочник, а лексикографическая версия национального языка и, таким образом, источник изучения языка. Именно так толковый словарь должен быть использован в процессе обучения, и лексикографическая интерпретация устаревшей лексики в толковых словарях должна давать представление о таких словах и выражениях как части лексического состава русского языка. С другой стороны, на современном этапе развития науки с учетом «особого места словарей в социокультурной ситуации конца XX — начала XXI в.», «отчетливой ориентации всех направлений современной лингвистики на обнародование результатов своей деятельности в словарной форме» (Козырев, Черняк 2014, 5) и комплексного характера современных словарей теоретические лингвистические задачи во многом решаются именно средствами лексикографии. В связи с этим и представление устаревшей лексики в толковых словарях современного русского языка вносит весомый вклад в решение не имеющего на настоящий момент однозначного ответа вопроса о статусе, классификации и особенностях функционирования этой группы слов и выражений.

Основные проблемы толковой лексикографии, связанные с представлением устаревшей лексики: критерии включения такой лексики в словник и принципы ее описания (система помет и типовые толкования, вопрос различения и разной интерпретации историзмов и архаизмов). Рассмотрим эти проблемы подробнее на примере лексикографических проектов, являющихся самыми новыми (т. е. наиболее близкими современному состоянию языка) и самыми подробными (толковыми словарями большого типа, т. е. многотомными изданиями, детально описывающими семантику и функционирование лексических единиц), а именно — на примере «Словаря современного русского языка», известного как БАС–3, и «Словаря русского языка XXI в.», создаваемого группой под руководством Г. Н. Скляревской.

Со сложностью определения того, что собой представляет устаревшая лексика, и связан вопрос включения тех или иных лексических единиц в словарь. Академические словари (и в частности, «Словарь современного русского языка» (БАС–3)) идут от «цитатного» материала, составляющего базу Словаря (таким образом, если слово зафиксировано в классической русской литературе, входящей в источниковую базу проекта, оно окажется в словаре с соответствующей пометой):

Взбутетенивать ... Устар. прост. Сильно избивать, колотить кого-л. *Долго преследовал парень побитый Барина бранью своей ядовитой: Мы-ста тебя взбутетеним дубьем Вместе с горластым твоим холуем!* Некр. Псов. охота....

В словник БАС–3 попадают даже слова без иллюстраций, зафиксированные в лексикографических источниках предыдущего периода:

Местоблюстителъ ... Устар. Тот, кто временно исполняет обязанности высшего духовного сановника.

А новаторский проект «Словарь русского языка XXI в.» под ред. Г. Н. Скляревской (Скляревская 2019), ориентирующийся на реальное речевое употребление, фиксирует лексику, используемую в живом общении, и с помощью типовых речений отмечает особенности ее современного использования: см.

Велеречие ... Устар. Красноречие, многословие. *Пышное в. § В соврем. употр. Ирон. В. и пустословие чиновников. Образец казенного велеречия. Удивляюсь, как слушатели не заснули от твоего велеречия. В. на совещании неуместно!*²

Такое внимательное описание функционирования устаревшей и устаревающей лексики в современном языке — важное достоинство «Словаря русского языка XXI в.». В ряде случаев в этом толковом словаре подчеркивается использование историзмов и архаизмов со специальными стилистическими целями (изображение соответствующей эпохи, стилизация языка под язык определенного времени): см.

Али, союз. Устар. и в стилизованной речи. Или (1, 4 и 5 зн.) (обычно в вопросительных предложениях). *Здесь пуда три али четыре будет. Вот не знаю, поймет он али нет. Ты пьяный али захворал? Идешь с нами али остаешься? Встретимся аль нет? Хочешь ярких камней аль цветочной парчи?* (Лермонтов);

Авиатрисса и авиатриса. *Ист.* В России в начале XX в.: женск. авиатор (2 зн.). *Первой авиатриссой стала французская баронесса де Ларош в 1910 году. Воспитанница Мариинского института благородных девиц Лидия Зверева считается первой русской авиатриссой. § В соврем. употр. (обычно в стилизованной речи). Очерки о современных авиатриссах. После войны ряды авиатрисс пополнила Марина Попович. А., повторившая подвиг Н. Гастелло.*

Также подробно описываются семантические нюансы употребления устаревшего слова в современном языке: см.

Аглицкий. *Устар.* Английский¹. *А. клуб. § В соврем. употр. Шутл. Аглицкая королева. А. шпион. А. язык;*

Байство. *Ист.* В Средней Азии досоветского периода: класс богачей, владевших крупными земельными, скотоводческими хозяйствами и имевших большую власть. *Киргизское б. Эксплуатация трудящихся феодальным байством. Ликвидация кулачества и байства. В Казахстане б. составляло примерно десять процентов всего населения. § В соврем. употр. Перен. О самоуправстве, деспотизме, попрании прав других людей. Критика политического кумовства и байства. Обвинение чиновника в байстве, произволе.*

Такая информация, как представляется, важна для читателя, так как помогает уяснить место и особенности употребления устаревшей лексики в словарном составе современного русского языка.

Что касается принципов описания языковых единиц минувших эпох, то в разных толковых словарях при их описании используются неодинаковые пометы (см. подробнее (Емельянова 2015)). БАС-3 различает *устар.* и *устаревающее* (ср. в «Словаре современного русского языка»: в 4-х томах под ред. А. Е. Евгеньевой (МАС) используется только помета *устар.*; о пометах в МАС — см. подробнее (Самотик 2015)). В словаре под редакцией Г. Н. Скляревской есть пометы *ист.*, *устар.* и *устаревающее*. Хронологические пометы при лексике, выходящей из активного употребления, во-первых, дают читателю информацию о степени и причинах ее архаизации (поэтому важно объяснить учащимся значение этих помет и обращать внимание на разные хронологические пометы в разных словарях), а во-вторых, могут помочь решению теоретического вопроса о группах (разрядах) исторической лексики в современном языке.

² Шрифтовое выделение помет, указывающих на функционирование в современном языке, наше.

На то, что слово относится к пласту устаревшей лексики, указывают не только пометы, но и элементы толкования (см. в *советское время*, в *Древней Руси* и т. п.). В этом отношении особенно сложно толкование советизмов, которые, с одной стороны, относятся к пласту не очень давно устаревшей лексики (и не всегда осознаются, в частности, обучающимися как таковые), а с другой стороны, могут быть привязаны к разным периодам существования СССР (в связи с чем включение в толкование элемента «*В советск. время...*» не всегда является достаточным и исчерпывающим). См., например, дезориентирующее читателя отсутствие такой пометы при слове микропора в БАС–3: *микропора* ‘Устар. Разг. Подошва из микропористой резины’ и дифференциацию толкований в «Словаре русского языка XXI в.»: *антисоветчик* ‘Ист. В советск. время: человек, настроенный враждебно по отношению к СССР, советскому строю, его идеологии; гражданин СССР, занимавшийся антисоветской деятельностью (и подвергавшийся за это гонениям и репрессиям)’ и *валютница* ‘Ист. В СССР в 1980–1990-е годы: валютная проститутка’. Такая часть толкования — это и очень важный инструмент лексикографа, позволяющий показать хронологическую неоднородность устаревшей лексики, и дополнительный аргумент для выделения разных групп лексики, вышедшей из употребления. Обобщение и анализ таких элементов толкования может способствовать изучению вопроса о единой или разной интерпретации лексики разных хронологических пластов.

Сложной остается проблема показа в словаре процесса актуализации лексики. В этом отношении интересна лексикографическая практика динамических словарей под редакцией Г. Н. Склярёвской: см. например, (ДАС 1998), где возвращение лексики в активный запас языка показано с помощью специальных значков и приемов. Слова и устойчивые выражения в этом словаре снабжаются не только лингвистическим, но и полезным энциклопедическим комментарием: см., например, «*закон Божий* (учебный предмет в некоторых средних школах, знакомящий учащихся с основами христианства). После революции 1917 года преподавание закона Божия в СССР было отменено, а после 1929 года уголовно наказуемо; в настоящее время возрождается».

В целом могут быть сделаны следующие выводы общекультурного, методологического, лингвистического теоретического и прикладного характера:

— при том, что устаревшие и устаревающие слова и выражения входят в пассивный состав языка, их необходимо понимать и правильно использовать, потому что это позволяет быть культурным человеком, знающим свой язык и историю своего народа;

— отвечающее современному уровню развития науки преподавание тем, связанных с устаревшей лексикой, и доступное пользователю описание этой лексики в современных словарях русского языка — важные предпосылки такого знания; в процессе изучения устаревшей лексики необходимо использовать словари редких и устаревших слов, исторические словари и обязательно — толковые словари для уяснения особенностей использования таких слов и выражений в современном языке;

— современный антропоцентрически ориентированный толковый словарь не просто фиксирует сохраняющуюся в пассивном составе языка устаревшую лексику, но с помощью метаязыка дает информацию читателю о причинах выхода этих слов и устойчивых выражений из активного состава, степени устаревания лексической единицы, а главное — о ее стилистическом потенциале, возможностях и особенностях использования такой лексики в современном языке;

— дискуссионные теоретические вопросы, связанные с исследованием устаревшей лексики, могут быть решены во многом средствами современной лексикографии.

Словари

- БАС–3 — Горбачевич, К. С. (ред.). (2004–) *Большой академический словарь русского языка*: в 30 т. Т. 1–25 (продолжающееся издание). СПб.: Наука.
- Бирюкова, С. К. (2003) *Словарь культуроведческой лексики русской классической литературы*. СПб.: Просвещение, 351 с.
- Глинкина, Л. А. (сост.). (1998) *Иллюстрированный словарь забытых и трудных слов из произведений русской литературы XVIII–XIX вв.* Оренбург: Книжное издательство, 276 с.
- ДАС — Скляревская, Г. Н. (ред.). (1998) *Толковый словарь русского языка конца XX века: языковые изменения*. СПб.: Фолио-Пресс, 700 с.
- Елистратов, В. С. (1997) *Язык старой Москвы: лингвоэнциклопедический словарь*. М.: Русские словари, 703 с.
- МАС — Евгеньева, А. П. (ред.). (1981–1984) *Словарь русского языка*: в 4 т. М.: Русский язык.
- Рогожникова, Р. П., Карская, Т. С. (1996) *Школьный словарь устаревших слов русского языка: по произведениям русских писателей XVIII–XX вв.* М.: Просвещение: Учебная литература, 608 с.
- Скляревская, Г. Н. (ред.). (2019) *Словарь русского языка XXI века*. СПб.: Изд-во РГПУ им. А. И. Герцена, 564 с.
- Сл18 — Сорокин, Ю. С. (ред.). (1984–) *Словарь русского языка XVIII в.* Т. 1–22 (продолжающееся издание). Л.–СПб.: Наука.
- СЛРЯ11-17 — *Словарь русского языка XI–XVII вв.* (1975–) Т. 1–30 (продолжающееся издание). М.: Наука.
- Сомов, В. П. (1996) *Словарь редких и забытых слов*. М.: Владос, 763 с.
- СОРЯ — Мжельская, О. С. (ред.). (2004–) *Словарь обиходного русского языка Московской Руси XVI–XVII вв.* Т. 1–8 (продолжающееся издание). СПб.: Наука.
- Ткаченко, Н. Г., Андреева, И. В., Баско, Н. В. (сост.). (1997) *Словарь устаревших слов: По произведениям школьной программы*. М.: Айрис-пресс, 257 с.
- Федосюк, Ю. А. (2000) *Что непонятно у классиков, или Энциклопедия русского быта XIX века*. 3-е изд. М.: Флинта, 263 с.

Литература

- Аркадьева, Т. Г., Васильева, М. И., Шарри, Т. Г. и др. (2014) Архаизмы и субстракты в русском языке. *Филологические науки. Вопросы теории и практики*, 10–2 (40): 24–29.
- Емельянова, О. Н. (2015) Устаревшая лексика в системе языка (по материалам толковых словарей современного русского языка). *Экология языка и коммуникативная практика*, 2 (5): 48–69.
- Зозикова, М. Е. (2015) О некоторых тенденциях употребления устаревшей лексики в современном русском языке. *Филология и литературоведение*, 2 (41): 34–40.
- Козырев, В. А., Черняк В. Д. (2014) Современные ориентации отечественной лексикографии. *Вопросы лексикографии*, 1 (5): 5–15.
- Норман, Б. Ю. (2016) Уходящее слово: историзм, архаизм, нотиолизм? *Коммуникативные исследования*, 4 (10): 21–38.
- Попов, Р. Н. (1995) Архаизмы в структуре современных фразеологических оборотов. *Русский язык в школе*, 3: 86–90.
- Правдина, И. С., Чуриков, С. А. (2016) Словарь асимметричных архаизмов русского языка. В кн.: *Всероссийский форум русского языка, посвященный наследию академика И. И. Срезневского, 12–15 апреля 2016 года*. Рязань: Рязанский государственный университет им. С. А. Есенина, с. 111–112.
- Самотик, Л. Г. (2015) Устаревшая лексика в толковых словарях русского языка. В кн.: И. В. Пекарская и др. (ред.). *Актуальные проблемы изучения языка, литературы и журналистики: история и современность: Материалы X Международной научно-практической конференции (24–26 ноября 2015 г., Абакан, Россия)*. Абакан: Изд-во ФГБОУ ВПО «Хакасский государственный университет им. Н. Ф. Катанова», с. 33–35.
- Шанский, Н. М. (1972) *Лексикология современного русского языка*. М.: Просвещение, 327 с.

Dictionaries

- BAS–3 — Gorbachevich, K. S. (ed.). (2004–) *Boʻshoj akademicheskij slovarʻ russkogo yazyka*: In 30 vols. Vol. 1–25 (continued ed.). Saint Petersburg: Nauka Publ. (In Russian)

- Biryukova, S. K. (2003) *Slovar' kul'turovedcheskoj leksiki russkoj klassicheskoj literatury*. Saint Petersburg: Prosveshchenie Publ., 351 p. (In Russian)
- DAS — Sklyarevskaya, G. N. (ed.). (1998) *Tolkovyj slovar' russkogo yazyka kontsa XX veka: yazykovye izmeneniya*. Saint Petersburg: Folio-Press Publ., 700 p. (In Russian)
- Elistratov, V. S. (1997) *Yazyk staroj Moskvy: Lingvoentsiklopedicheskij slovar'*. Moscow: Russkie slovari Publ., 703 p. (In Russian)
- Fedosyuk, Yu. A. (2000) *Chto neponyatno u klassikov, ili Entsiklopediya russkogo byta XIX veka*. 3rd ed. Moscow: Flinta Publ., 263 p. (In Russian)
- Glinkina, L. A. (comp.). (1998) *Ilyustrirovannyj slovar' zabytykh i trudnykh slov iz proizvedenij russkoj literatury XVIII–XIX vv.* Orenburg: Knizhnoe izdatel'stvo Publ., 276 p. (In Russian)
- MAS — Evgen'eva, A. P. (ed.). (1981–1984) *Slovar' russkogo yazyka*: In 4 vols. Moscow: Russkij yazyk Publ.
- Rogozhnikova, R. P., Karskaya, T. S. (1996) *Shkol'nyj slovar' ustarevshikh slov russkogo yazyka: Po proizvedeniyam russkikh pisatelej XVIII–XX vv.* Moscow: Prosveshchenie Publ.: Uchebnaya literatura Publ., 608 p. (In Russian)
- Sklyarevskaya, G. N. (ed.). (2019) *Slovar' russkogo yazyka XXI veka*. Saint Petersburg: Herzen State Pedagogical University of Russia Publ., 564 p. (In Russian)
- Sl18 — Sorokin, Yu. S. (ed.). (1984–) *Slovar' russkogo yazyka XVIII v.* Vol. 1–22 (continued ed.). Leningrad–Saint Petersburg: Nauka Publ. (In Russian)
- SLRYa11–17 — *Slovar' russkogo yazyka XI–XVII vv.* (1975–) Vol. 1–30 (continued ed.). Moscow: Nauka Publ. (In Russian)
- Somov, V. P. (1996) *Slovar' redkikh i zabytykh slov*. Moscow: Vldos Publ., 763 p. (In Russian)
- SORYa — Mzhel'skaya, O. S. (ed.). (2004–) *Slovar' obikhnodnogo russkogo yazyka Moskovskoj Rusi XVI–XVII vv.* Vol. 1–8 (continued ed.). Saint Petersburg: Nauka Publ. (In Russian)
- Tkachenko, N. G., Andreeva, I. V., Basko, N. V. (comp.). (1997) *Slovar' ustarevshikh slov: Po proizvedeniyam shkol'noj programmy*. Moscow: Ajris-press Publ., 257 p. (In Russian)

References

- Arkad'eva, T. G., Vasil'eva, M. I., Sharri, T. G. et al. (2014) Arkhaizmy i substrakty v russkom yazyke [Archaisms and substrates in the Russian language]. *Filologicheskie nauki. Voprosy teorii i praktiki*, 40 (10–2): 24–29. (In Russian)
- Emelyanova, O. N. (2015) Ustarevshaya leksika v sisteme yazyka (po materialam tolkovykh slovarej sovremennogo russkogo yazyka) [Archaisms in the system of the Russian language (on the material of Russian explanatory dictionaries)]. *Ekologiya yazyka i kommunikativnaya praktika — Ecology of Language and Communicative Practice*, 2 (5): 48–69. (In Russian)
- Kozyrev, V. A., Chernyak, V. D. (2014) Sovremennye orientatsii otechestvennoj leksikografii [Modern orientations of Russian lexicography]. *Voprosy leksikografii — Russian Journal of Lexicography*, 1 (5): 5–15. (In Russian)
- Norman, B. Yu. (2016) Ukhodyashchee slovo: istorizm, arkhazim, notiolizm? [A disappearing word: Historicism, archaism, notiolism?]. *Kommunikativnye issledovaniya — Communication Studies*, 4 (10): 21–38. (In Russian)
- Popov, R. N. (1995) Arkhaizmy v strukture sovremennykh frazeologicheskikh oborotov. *Russkij yazyk v shkole — Russian Language at School*, 3: 86–90. (In Russian)
- Pravdina, I. S., Churikov, S. A. (2016) Slovar' asimmetrichnykh arkhazimov russkogo yazyka. In: *Vserossijskij forum russkogo yazyka, posvyashchennyj naslediyu akademika I. I. Sreznevskogo, 12–15 aprelya 2016 goda*. Ryazan: Ryazan State University named for S. Yesenin Publ., pp. 111–112. (In Russian)
- Samotik, L. G. (2015) Ustarevshaya leksika v tolkovykh slovaryakh russkogo yazyka [Vocabulary in the explanatory dictionaries of the Russian language]. In: I. V. Pekarskaya et al. (eds.). *Aktual'nye problemy izucheniya yazyka, literatury i zhurnalistiki: istoriya i sovremennost'*: Materialy X Mezhdunarodnoj nauchno-prakticheskoy konferentsii (24–26 noyabrya 2015 g., Abakan, Rossiya). Abakan: Khakassian State University named after N. F. Katanov Publ., p. 33–35. (In Russian)
- Shanskij, N. M. (1972) *Leksikologiya sovremennogo russkogo yazyka*. Moscow: Prosveshchenie Publ., 327 p. (In Russian)
- Zozikova, M. E. (2015) O nekotorykh tendentsiyakh upotrebleniya ustarevshej leksiki v sovremennom russkom yazyke [About some trends in the use of obsolete lexis in the contemporary Russian language]. *Filologiya i literaturovedenie — Philology and Literature*, 2 (41): 34–40. (In Russian)

Сведения об авторе:

Елена Владимировна Генералова, e-mail: elena-generalova@yandex.ru

Для цитирования: Генералова, Е. В. (2019) Устаревшая лексика русского языка: вопросы преподавания и лексикографической интерпретации. *Journal of Applied Linguistics and Lexicography*, 1 (2): 371–380. DOI: 10.33910/2687-0215-2019-1-2-371-380

Получена 14 августа 2019; прошла рецензирование 11 сентября 2019; принята 12 сентября 2019.

Финансирование: Работа выполнена при поддержке гранта РФФИ № 20-012-00122.

Права: © Автор (2019). Опубликовано Российским государственным педагогическим университетом им. А. И. Герцена. Открытый доступ на условиях лицензии CC BY-NC 4.0.

Author:

Elena V. Generalova, e-mail: elena-generalova@yandex.ru

For citation: Generalova, E. V. (2019) Outdated vocabulary of the Russian language: questions of teaching and lexicographic interpretation. *Journal of Applied Linguistics and Lexicography*, 1 (2): 371–380. DOI: 10.33910/2687-0215-2019-1-2-371-380

Received 14 August 2019; reviewed 11 September 2019; accepted 12 September 2019.

Funding: The research was funded by Russian Foundation for Basic Research, grant no. 20-012-00122.

Copyright: © The Author (2019). Published by Herzen State Pedagogical University of Russia. Open access under CC BY-NC License 4.0.

ИСТОРИИ, МУЗЕИ, СЛОВАРИ: ФОРМЫ ПРЕДСТАВЛЕНИЯ ЗНАНИЯ О ЯЗЫКАХ В XVI ВЕКЕ И ПЕРВЫЕ ЛИНГВИСТИЧЕСКИЕ СПРАВОЧНИКИ (КНИГИ-ПОЛИГЛОТЫ)¹

М. Л. Сергеев✉¹

¹ Российский государственный педагогический университет им. А. И. Герцена, 191186, Россия,
Санкт-Петербург, наб. реки Мойки, д. 48

HISTORIES, MUSEUMS AND DICTIONARIES: THE FORMS OF REPRESENTING LINGUISTIC KNOWLEDGE IN THE 16TH CENTURY AND THE EARLIEST LINGUISTIC HANDBOOKS

M. L. Sergeev✉¹

¹ Herzen State Pedagogical University of Russia, 48 Moika River Emb., Saint Petersburg 191186, Russia

Аннотация. В статье рассматриваются способы представления сведений о многообразии языков мира в XVI веке — в период стремительного увеличения объема лингвистической информации (в том числе о языках Востока и Нового Света), доступной европейским ученым. Отсутствие науки о языках в качестве самостоятельной научной дисциплины в то время не означало отсутствия лингвистического интереса. Однако собираемые авторами сведения о языках включались в сочинения, относившиеся к различным жанрам и областям знания: исследования по истории и географии, справочники по естественным наукам, коллекции алфавитов и переводов молитвы «Отче наш», многоязычные словари и т. д. Соединение и сопоставление разнородной информации о языках стали возможны благодаря появлению особого жанра ученой литературы — «многоязычных книг» (полиглотов), которые по сути были первыми лингвистическими справочниками, собравшими образцы различных языков и сведения по их истории (наиболее известны книги-полиглоты Г. Постеля, Т. Библиандера, К. Гесснера, А. Рокки и К. Дюре). Компилятивный характер этих сочинений предопределил сохранение в их тексте жанрового многообразия цитируемых источников. Согласно классификации, представленной в статье, основными формами представления лингвистического знания в XVI в. следует считать: (1) историю о происхождении народов

Abstract. In the history of humanities the 16th century was characterised by a considerable increase in the amount of information on world languages in absence of linguistics as an institutionalised scientific discipline. The latter did not mean that European scholars were not interested in linguistic issues; yet linguistic data collected by them was included in volumes that represented a variety of genres and fields of knowledge, such as treatises on history and geography, handbooks of natural sciences, collections of alphabets and translations of the Lord's Prayer, multilingual dictionaries, etc. The compilation and comparison of heterogeneous linguistic evidence collected from printed books and manuscripts, academic correspondence, and oral sources became possible due to the emergence of a new genre of scholastic literature — the multilingual books (or the polyglot-books), which consisted of language samples and short accounts of language history (e. g. the works by G. Postel, Th. Bibliander, C. Gessner, A. Rocca and C. Duret). Polyglot-books were, to a large extent, compilations of quotations, which preserved the formal diversity of their sources. The author proposes a classification of the main forms of presenting linguistic knowledge, namely (1) a history of a nation and its language; (2) a multilingual (comparative) dictionary; and (3) a collection of language samples. It is suggested (with a special reference to the «Mithridates» (1555) by C. Gessner) that these forms must have influenced the structure of the polyglot-books, thus allowing them to become a tool for solving several

¹ В основу статьи положены материалы докладов, прочитанных на конференциях «Internationaler Kongress Conrad Gessner (1516–1565)» (Цюрих, Universität Zürich, 6–9 июня 2016 г.) и «В тени Просвещения: схоласты, гуманисты и эрудиты на пороге Нового времени» (Москва, НИУ ВШЭ, 17–18 сентября 2018 г.)

и языков, (2) многоязычный (сопоставительный) словарь и (3) собрание языковых образцов. На примере справочника Конрада Гесснера «Митридат. О различиях языков» (1555) показано возможное влияние этих форм на структуру книг-полиглотов, которое позволило им стать инструментом для решения ряда лингвистических задач, таких как упорядочение языковой номенклатуры и классификация языков и диалектов.

Ключевые слова: история лингвистики, XVI век, словари, справочники, история языка, классификация языков.

specific linguistic issues, such as multiplicity of language names and classification of languages and dialects.

Keywords: history of linguistics, 16th century, dictionaries, handbooks, language history, language classification.

Для истории языкознания и гуманитарных наук в целом XVI век представляет значительный интерес в связи с появлением в этот период первых заметных попыток систематизации накопленных знаний о языках мира, их сопоставления и классификации. Разумеется, интерес к истории языков, неразрывно связанной с историей народов, говоривших на них, имелся и у средневековых авторов (Bonfante 1954), а вопросы о происхождении и диалектном многообразии народного языка рассматривались в Италии уже в XIV–XV вв. (Степанова 2000). Однако проблема изучения и классификации всех языков мира, на фоне стремительного расширения языкового горизонта и накопления знаний о языках, была впервые поставлена, по-видимому, именно в XVI в., и решалась она преимущественно учеными, жившими к северу от Альп (ср. Van Hal, Considine 2010).

Для ученых-гуманистов того времени изучение языкового многообразия (как и занятия в любой другой области науки) предполагало в первую очередь ревизию знания, представленного в письменных источниках, от античности и книг Ветхого Завета до сочинений авторов-современников. Вместе с тем проводились сбор (или коллекционирование), публикация и анализ эмпирического материала древних и новых языков, не объединенные еще некоей единой исследовательской программой, но обусловленные весьма разнообразными мотивами научной деятельности, сосуществование которых создало благоприятную среду для появления науки о языках (см.: Van Hal 2010; Van Hal 2013).

Важнейшим мотивом изучения древних ближневосточных языков и греческого в XVI в. было толкование Священного Писания, чтение его оригинала (на древнееврейском и древнегреческом) и ранних переводов (сирийского, эфиопского, арабского), исследование семантики и этимологии отдельных терминов и имен (см.: Bobzin 2000, Kessler-Mesguich 2000). Сопоставление лексики ближневосточных языков достаточно рано привело к обнаружению родственных связей между ними — в пределах того, что позже будет названо семитской семьей языков (Kessler-Mesguich 2013, 24–26).

Библейская экзегеза была тесно связана с занятиями гуманистов, читавших античные первоисточники на языках оригинала (ср. Hamilton 1996). Внимательное изучение античного наследия дало ученым раннего Нового времени, с одной стороны, образец сопоставления двух тесно связанных языков — греческого и латыни, — предполагавшего решение вопроса об их относительной древности и генетической связи между ними (см. Tavoni 1986), а с другой стороны, модель описания диалектного многообразия (ср. Van Rooy 2017, 51–103); и то, и другое было вскоре применено к материалу народных языков.

Гуманистический интерес к греческой и латинской лексикологии оказался также исключительно важен для развития естественных наук, теоретическую основу которых продолжали составлять работы греческих и римских авторов. Античные названия животных, растений и минералов, описания их форм и медицинских свойств соотносились с данными

о природном разнообразии современного мира и естественнонаучной номенклатурой в живых языках — европейских и «экзотических». Это требовало выполнения разного рода филологических задач: составления многоязычных словарей, исключения слов-«призраков», возникавших в источниках по ошибке переписчиков, объяснения этимологии непонятных терминов, которая могла бы как-то прояснить их денотат (ср.: Ogilvie 2006: 87–138; Воробьев 2015; Vorobyev 2018).

Кроме того, издание и комментирование античных сочинений по истории позволило дополнить традиционную библейскую картину «смешения» языков и расселения потомков Ноя древними свидетельствами о «варварских» народах, к которым возводили себя европейцы XVI в. Содержащиеся в этих текстах лингвистические наблюдения и глоссы вызвали живой интерес и многократно интерпретировались, так как предполагаемое «сходство» языков использовалось в качестве исторического аргумента при доказательстве родства тех или иных народов или установлении мест их первоначального расселения (ср.: Margolin 1985; Сергеев 2018, 36–52).

Для знакомства с живыми неевропейскими языками, на которых говорили в Юго-Восточной Азии и Новом Свете, решающее значение имела деятельность миссионеров, составлявших двуязычные грамматики и словари для нужд миссии и нередко включавших образцы перевода религиозных текстов на эти языки в своих сочинения и письма (ср. Demonet, Uetani 2008). Используя эти источники, К. Дюре при составлении «Сокровищницы истории языков мира» (1613) уже имел возможность сравнить египетские иероглифы с письменностью китайцев и американских индейцев (Duret 1619, 378–389); кроме главы, посвященной «языку восточных индейцев в общем», у него есть отдельные главы о китайском и японском (Duret 1619, 883–922; ср. Simon 2011, 225–239).

Совместное или разрозненное действие этих мотивов в сочетании с характерным для европейцев того времени лингвистическим любопытством (выходившим за пределы компетенций, необходимых для решения практических задач: ср. Considine 2017, 11–30) привело к накоплению значительного лингвистического материала, в том или ином виде представленного в ученой литературе. Сведения о языках в текстах могли быть представлены в трех основных формах, которые как будто соотносятся с тремя заметными вопросами в лингвистической мысли XVI–XVII вв. — о происхождении, сходствах и различиях и многообразии языков:

1) Рассказ о происхождении народов и языков, содержащий сведения о генеалогии, географии, этнографии, религии, а также этимологии и диалектологии, письменности и литературе, которые приводились для подкрепления исторических взглядов автора; такое сочетание разнообразных сведений мы обнаруживаем в сочинениях о германских древностях, вдохновленных проектом К. Цельтиса *Germania illustrata*: «Описании Германии» (1518) Ф. Иреника, «Анналах баварских князей» (1522) И. Авентина, «Трех книгах германской истории» (1531) Б. Ренана, позднее — в «Древней Германии» (1616) Ф. Клювера (ср.: McLean 2007, 105–115; Mundt 2008).

2) Многоязычные словари, а также списки слов на разных языках, как правило демонстрирующие сходство и родство между ними. Среди составителей алфавитных словарей наибольшего охвата языков удалось достичь И. Мегизеру (1553–1619) — в «Многоязычном тезаурусе» («*Thesaurus polyglottus*») представлен (разумеется, неравномерно) лексический материал не только европейских и ближневосточных языков, но и малайского, японского, языков Кубы, Гаити, Флориды и др. (см. перечень языков и диалектов: Megiser 1603, (:)За-(:)ба; ср. Alston, Danielsson 1964). Значительное распространение получили лексические сопоставления в форме списков: их образцы сохранились в составе исторических («Альпийская Реция» (1538) Э. Чуди, «О переселениях народов» (1557)

В. Лация) и лингвистических сочинений («Древнееврейский словарь» (1523) С. Мюнстера, «Трактат о сходстве французского и греческого языков» (1565) А. Этьенна), а также в текстах писем (Ю. Липсия, Ф. Сассетти).

3) Коллекции языковых образцов разного типа: в зависимости от интересов составителя и доступности источников в этой роли могли выступать отдельные слова, тематические группы слов, фразы и тексты, алфавиты, образцы письма. Так, например, в пособии по каллиграфии У. Висса после подробного разбора разновидностей латинского письма приведены также другие «алфавиты» («*varia Alphabeta*»): еврейский, греческий, эфиопский, сирийский и др. (Wyss 1549, N1a–O2a). Еще большее их разнообразие — включающее арабский, армянский, славянский глаголический, японский и т. д. — мы находим в «Трактате о шифрах и тайнописи» (1586) Б. де Виженера (ср. Simon 2018, 311–317). Некоторые переводы молитвы «Отче наш» — на шведский, финский и латышский — собраны С. Мюнстером в «Космографии» (Münster 1550, 789, 847), а перечень наиболее употребительных слов местного бразильского языка был включен натуралистом Г. Маркграфом в 8-ю книгу «Естественной истории Бразилии» (Piso, Markgraf 1648, 276–277 2-й паг.).

Границы дискурсов затрудняли последовательное изучение и критику свидетельств о языках и их трактовок, содержащихся в литературе, тем более что лингвистика еще не имела статуса самостоятельной научной дисциплины, и проблемы истории и родства языков не существовало в чистом виде. Однако развитие научного книгопечатания и технологий организации информации (Blair 2010, 62–172) благоприятствовало появлению особого рода компилятивных справочных изданий, аккумулировавших разнородные свидетельства о языках, — так называемых *полиглотов* (буквально — книг, *говорящих на множестве языков*).¹ В них сообщались сведения об истории языков и народов, говоривших на них, рассматривалось происхождение алфавитов, приводились лексические сопоставления и, как правило, публиковались образцы текстов на разных языках. К первым *полиглотам* относятся: «Алфавит 12 языков, различающихся буквами» («*Linguarum duodecim characteribus differentium alphabetum*», 1538) Г. Постеля, «Комментарий об общем принципе всех языков и письменностей» («*De ratione communi omnium linguarum et literarum commentarius*», 1548) Т. Библиандера, «Митридат. О различиях языков» («*Mithridates De differentiis linguarum*», 1555) К. Гесснера, «Приложение о диалектах, то есть о различных родах языков» («*Appendix de dialectis, hoc est de variis linguarum generibus*», 1591) А. Рокки, «Сокровищница истории языков этого мира» («*Thresor de l'histoire des langues de cest univers*», 1613) К. Дюре и т. д.²

Следует заметить, что справочники-полиглоты были весьма неоднородны по содержанию и номенклатуре представленных в них языков. Авторы делали акцент на разных темах и, как правило, уделяли более пристальное внимание родному языку: так, Г. Постель последовательно приводит алфавиты и образцы текстов в оригинальной графике, а Т. Библиандер одним из первых рассмотрел сходства в грамматической структуре языков (ср. Peters 1984); швейцарец К. Гесснер особенно подробно освещает историю немецкого языка (Gessner 1555, 27a–44b), итальянец А. Рокка — латинского и итальянского (Rosca 1591, 342–351; ср. Fiacchi 1996) француз К. Дюре не включает главу о французском в свой справочник, обещая написать о нем отдельно (Duret 1619, 867) (впрочем, эти планы не осуществились).

Изучение источников и композиции этих справочников показывает, что значительную часть текста «полиглотов» составляют цитаты, содержащие суждения о языках и языковые

¹ Ср. использование этого термина для характеристики трактата «Митридат. О различиях языков» в одном из писем К. Гесснера: «*Mithridates meus πολυγλωττος*» (Gessner 1577, 26b).

² Другие примеры книг-полиглотов упоминают: (Law 2003, 218–223, Swiggers 1997, 139–140, Simon 2018).

образцы: они часто приводились дословно и, как правило, сохраняли некоторую содержательную и формальную обособленность, не подчиняясь единому «голосу» повествования. Таким образом, читателям справочников-полиглотов становились доступны тексты о языках, написанные первоначально с разными целями и включенные в работы разных жанров и различной тематики (словари, грамматики, книги по истории, справочники по географии, зоологии и ботанике и т. д.). Вместе с тем представляется, что разнородность источников влияла на структуру компиляций, требуя от составителя искать подходящие формы упорядочения информации. Этот тезис будет раскрыт далее на примере «Митридата» (1555) К. Гесснера — компактного алфавитного справочника, включившего информацию о более чем 100 языках и диалектах.³ «Митридат» был хорошо известен современникам Гесснера и в последовавшие десятилетия был дважды переиздан с существенными дополнениями: в 1591 г., без упоминания имени Гесснера, под редакцией А. Рокки (Росса 1591), и в 1610 г. — с подробными комментариями К. Вазера (Waser 1610).

Научные занятия Конрада Гесснера (1516–1656) были чрезвычайно разнообразны и принесли значительные плоды в разных областях знания: он составил первую универсальную библиографию («*Bibliotheca universalis*», 1545–1555) и первую энциклопедию по зоологии («*Historia animalium*», 1551–1587), справочники по медицине и минералогии, собирал материалы для «Истории растений», которые, к сожалению, не успел издать. Вместе с тем, начиная со студенческих лет, он участвовал в составлении и редактировании словарей, изданиях и переводах на латынь греческих авторов (см. Сергеев 2018, 87–110). Эти занятия несомненно повлияли на тематические акценты «Митридата» и отбор источников для него.

Изучение содержания и композиции отдельных глав «Митридата» и справочника в целом позволяет выделить в нем следующие способы представления лингвистической информации, в целом соответствующие рассмотренным выше формам:

1) *История языка* наиболее последовательно и подробно рассматривается в главах о германских языках (немецком, английском и «древнем галльском», который Гесснер считал диалектом германского), занимающих около трети от общего объема книги; глава о немецком включает также наблюдения о фламандском, исландском, готском и норвежском. В представлениях автора, немецкий язык по своей древности и богатству словаря не уступал языкам классическим, находясь при этом в родстве с греческим языком (Gessner 1555, 34b–35b; Maaler 1561, *3b). Древность немецкого языка подтверждалась, например, предполагаемыми германскими этимонами для отдельных терминов в сочинениях римских авторов, сопоставлением галльских и немецких двусоставных имен на -bald(us), -man(us), -rich (-rix) (Gessner 1555, 18a–19a, 32b–34b), а также обнаружением созвучных слов во французском и немецком (вроде *burgois* ~ *burger* ‘горожанин’, *banche* ~ *banck* ‘скамья’, *cuquelin* ~ *kuechlin* ‘пирожок’, *cuysin* ~ *kussin* ‘подушка’, *tailler* ~ *teller* ‘тарелка’ и т. д.⁴), которые немецкие авторы охотно объясняли как реликты галльского субстрата, отождествляемого с германским языком (Gessner 1555, 20a, 38a/b). Текст этих глав составляет достаточно связанное изложение, в котором рассмотрение языковых особенностей предваряется и дополняется историческими и этнографическими сведениями о древних германских и кельтских племенах. Основными источниками этой информации стали сочинения историков — Цезаря, Тацита, Аммиана Марцеллина, Эйнхарда, И. Авентина, И. Вадяна и др., из которых нередко даются пространные цитаты (ср. Сергеев 2018, 44–52, 169–182).

³ Об истории создания «Митридата» и филологических занятиях Гесснера см.: (Сергеев 2018, 59–196; Sergeev 2019); см. также обстоятельно прокомментированное издание справочника, подготовленное Б. Коломба и М. Петерсом (Gessner, Colombat, Peters 2009). Отдельные главы «Митридата» были недавно переведены на русский язык М. В. Шумилиным (Шумилин 2016).

⁴ Некоторые из приведенных Гесснером соответствий действительно объясняются значительной долей германизмов во французском словаре.

2) *Языковые образцы* (*specimina linguarum*) имеются в большинстве глав «Митридата» (кроме отсылочных статей) и весьма различаются по характеру и объему: от отдельных глосс (например, «*nabis* на эфиопском значит ‘жираф’», «*parasanga, schoenus* и *astarus* — персидские названия мер», «Ἀρράβαξ — ‘танцор’ или ‘хулитель’ на фракийском языке»: Gessner 1555, 6a, 63b, 69a) до списков слов и целых текстов. Одним из источников этих образцов были корреспонденты Гесснера, другие заимствованы из книг, в том числе — справочников и словарей, в составлении или издании которых участвовал Гесснер: например, названия животных на египетском (Gessner 1555, 5a) взяты из «*Historia animalium*»; трижды в справочнике цитируется греческий словарь Фаворино (Gessner 1555, 60a/b, 67b, 69a). Регулярно в качестве языкового образца приводятся переводы Господней молитвы (в «Митридате» процитированы 27 версий на различных языках и диалектах), которая стала основным образцом в позднейших книгах-полиглотах.⁵ Большинство переводов (на 22 языках) были изданы в том же (1555) году в виде отдельной таблицы, в которой тексты расположены в соответствии с представлениями Гесснера об истории и родстве языков: в первом столбце приведены тексты на древнееврейском и других семитских языках, затем на греческом и латинском, далее отдельными группами следуют переводы на романские, славянские и германские языки. Все эти тексты, глоссы и списки слов (названия месяцев, числительных) были экспонатами языковой коллекции Гесснера, которую он собирал наряду с коллекциями книг, растений, животных и минералов. Проводя аналогию с бумажными музеями — собраниями изображенных натуралий и артифициалий (ср. Meijers 2005) — можно говорить о «Митридате» как о музее языков.

3) *Словарная форма* присутствует в структуре справочника в двух качествах. Во-первых, как способ подачи языкового материала: таблицы лексических сопоставлений приведены в главах об армянском, эфиопском и — более подробная — в разделе о швейцарском диалекте немецкого (Gessner 1555, 10a/b, 7b, 38a/b). Также в виде приложения к тексту и одновременно в качестве развернутого языкового образца в книге опубликован словарь тайного языка *Rotwelsch* (Gessner 1555, 73b–77b). Во-вторых, словарный принцип использован для организации информации в справочнике в целом: ключевыми словами, определяющими деление текста на главы, выступают глоттонимы или этнонимы (названия народов-носителей языков), расположенные по алфавиту. В названиях глав проявился также интерес автора к синонимии и омонимии внутри этой группы терминов, например: «*De Chaldaica lingua, quae et Aramaea, & Syrogum, & Assyriogum, & Babyloniorum vocatur*» (‘О халдейском языке, который также называется арамейским, и языком сирийцев, и [языком] ассирийцев, и [языком] вавилонян’), «*De Illyrica sive Sarmatica lingua*» (‘Об иллирийском, или сарматском языке’); «*De Gallica lingua vetere*» / «*De Gallica lingua recentiore*» (‘О древнем галльском языке’ / ‘О современном галльском языке’ [то есть французском. — М. С.]), «*De Graeca lingua vetere*» / «*De lingua Graeca vulgari hodie*» (‘О древнем греческом языке’ / ‘О нынешнем народном греческом языке’) (Gessner 1555, 15a, 52a, 17b, 25b, 44b, 46b). Кроме того, в справочнике учтены словообразовательные и орфографические варианты латинских этнонимов, например: «*Bessi, hodie Bosnenses vel Bosnasienses*» (боснийцы), «*Croati alias Chroati*» (хорваты), «*Huni vel Hunni*» (гунны), «*Prusiae incolae & Pruteni dicuntur*» (пруссы) (Gessner 1555, 12b, 15b, 52a, 64b), и т. д. В качестве словаря глоттонимов «Митридат» дополняет изданный Гесснером в 1544 году латинский «Ономастикон», включавший антропонимы, этнонимы, топонимы, но не названия языков.

⁵ Вплоть до XIX в., ср. издание: Marcel J. J. (ed.). (1805) *Oratio Dominica CL linguis versa. Et propriis cuiusque linguae caracteribus plerumque expressa*. Parisiis, 150 fol.

Историки языкознания, давая оценку «Митридату», обращали преимущественное внимание на главы о германских языках, образующие связный исторический нарратив.⁶ Между тем при составлении справочника автор в равной мере использовал и другие формы представления знаний о языках: словарь (ономастикон) и музей. Выбор словарной формы кажется совершенно естественным для Гесснера, который видел в словаре универсальный инструмент для сохранения и поиска разнородной, не всегда поддающейся интерпретации информации: в условиях недостатка и постоянного пополнения знаний о языках мира, а также неопределенности в употреблении старых и новых глоттонимов обращение к словарному жанру кажется особенно удачным. Напротив, форма музея, наиболее отчетливо явленная в приложении к «Митридату», позволяла преодолеть важнейший недостаток алфавитной организации — неиерархизованность информации. В таблице языковых образцов визуально показана классификация языков (соответствующая расположению переводов «Отче наш») и вместе с тем непосредственно дано важное основание такой классификации — сходство и различие эмпирического языкового материала (для большей наглядности все переводы даны в латинской транскрипции). Впрочем, в таблице оказалась представлена лишь часть языков — те, для которых удалось получить необходимый образец текста.

Проект языкового справочника отнюдь не представлялся Гесснеру завершенным с выходом «Митридата» в 1555 г. В эпилоге он пишет о возможности нового издания: «Я осмелился написать обо всех языках <...> чтобы побудить других написать об отдельных или о нескольких языках, или в собственных сочинениях (что я предпочел бы), или мне — для пополнения или исправления в будущем нашего труда» (Gessner 1555, 78a). Для продолжения работы было создано хорошее основание: как видно из нашего обзора, структура «Митридата» была открыта для включения и упорядочения новой информации самого разного рода — как о тех языках, о которых известно только название, так и о тех, которым может быть посвящена отдельная история.

Sources

- Duret, C. (1619) *Thresor de l'histoire des langues de cest univers*. 2nd ed. Yverdon: De l'Imprimerie de la Societé Helvetiale Caldoresque, [32], 1030 p. (In French)
- Gessner, C. (1555) *Mithridates. De differentiis linguarum tum veterum tum quae hodie apud diversas nationes in toto orbe terrarum in usu sunt. Tigurini observationes*. Tiguri: Excudebat Froschoverus, [2], 78 fol. (In Latin)
- Gessner, C. (1577) *Epistolarum medicinalium, Conradi Gesneri, philosophi et medici Tigurini, libri III*. Tiguri: Excudebat Christoph. Frosch., [8], 140, 28 fol. DOI: 10.5962/bhl.title.151756 (In Latin)
- Maaler, J. (1561) *Die teütsch Sprach: alle Wörter, Namen und Arten zuo reden in hochteütscher Sprach, dem ABC nach ordenlich gestellt unnd mit guotem Latein gantz fleissig unnd eigentlich vertolmetscht...* Tiguri: Excudebat Christophorus Froschoverus, [8], 536 fol. (In German)
- Megiser, H. (1603) *Thesaurus polyglottus, vel dictionarium multilingue: ex quadringentis circiter tam veteris, quam novi ... orbis nationum linguis, dialectis, idiomatibus et idiotismis constans*: In 2 vols. Francofurti ad Moenum: Sumptibus Authoris. (In Latin)
- Münster, S. (1550) *Cosmographiae universalis Lib. VI*. Basileae: H. Petri, [24], 1162, [6] p. (In Latin)
- Piso, W., Markgraf, G. (1648) *Historia naturalis Brasiliae*. Lugduni Batavorum; Amstelodami: F. Hackius; L. Elzevirus, [12], 122, [10], 293, [8] p. (In Latin)
- Waser, C. (1610) *Mithridates Gesneri, exprimens differentias linguarum, tum veterum, tum quae hodie, per totum terrarum orbem, in usu sunt*. Ed. altera. Tiguri: Typis Vvolphianis, 140 fol. (In Latin)
- Wyss, U. (1549) *Libellus valde doctus, elegans, et utilis, multa et varia scribendarum literarum genera complectens*. Tiguri: per Urb. Wyses, [58] fol.

⁶ См. библиографию: (Сергеев 2018, 60).

References

- Alston, R. C., Danielsson, B. (1964) The earliest dictionary of the known languages of the world. *English Studies*, 45 (1–6): 9–13. DOI: 10.1080/00138386408597178 (In English)
- Blair, A. (2010) *Too much to know: Managing scholarly information before the Modern Age*. New Haven, CT: Yale University Press, XV, 397 p. (In English)
- Bobzin, H. (2000) Der Unterricht des Hebräischen, Arabischen und anderer semitischer Sprachen sowie des Persischen und Türkischen in Europa (bis zum Ende des 18. Jahrhunderts). In: S. Auroux, E. F. K. Koerner (eds.). *Geschichte der Sprachwissenschaften*. Vol. 1. Berlin; New York: Mouton de Gruyter, pp. 728–734. (In German)
- Bonfante, G. (1954) Ideas on the kinship of the European languages from 1200 to 1800. *Cahiers d'histoire mondiale*, 1 (3): 679–699. (In English)
- Considine, J. (2017) *Small dictionaries and curiosity: Lexicography and fieldwork in Post-Medieval Europe*. Oxford: Oxford University Press, 336 p. (In English)
- Demonet, M.-L., Uetani, T. (2008) Les langues des Indes orientales entre Renaissance et Âge classique. *Histoire Épistémologie Langage*, 30 (2): 113–139. (In French)
- Fiacchi, C. (1996) Il “De dialectis” di Angelo Rocca e il “Mithridates” di Conrad Gesner. In: M. Tavoni (ed.). *Italia ed Europa nella linguistica del Rinascimento*. Vol. 2. Modena: Panini, pp. 333–341. (In Italian)
- Gessner, C.; Colombat, B., Peters, M. (eds.). (2009) *Mithridate. Mithridates (1555)*. Genève: Librairie Droz, 407 p. (In Latin and French)
- Hamilton, A. (1996) Humanists and the Bible. In: J. Kraye (ed.). *The Cambridge companion to Renaissance humanism*. Cambridge: Cambridge University Press, pp. 100–117. (In English)
- Kessler-Mesguich, S. (2000) L'étude de l'hébreu et les autres langues orientales à l'époque de l'humanisme. In: S. Auroux, E. F. K. Koerner (eds.). *Geschichte der Sprachwissenschaften*. Vol. 1. Berlin; New York: Mouton de Gruyter Press, pp. 673–680. (In French)
- Kessler-Mesguich, S. (2013) *Les études hébraïques en France: de François Tissard à Richard Simon (1508–1680)*. Genève: Librairie Droz Publ., XIV, 312 p. (In French)
- Law, V. (2003) *The history of linguistics in Europe: From Plato to 1600*. Cambridge: Cambridge University Press, XVII, 307 p. (In English)
- Margolin, J.-Cl. (1985) Science et nationalisme linguistique ou la bataille pour l'étymologie au XVIe siècle. Bovelles et sa postérité critique. In: *The fairest flower: The emergence of linguistic national consciousness in Renaissance Europe*. Firenze: Accademia della Crusca, pp. 139–165. (In French)
- McLean, M. (2007) *The Cosmographia of Sebastian Münster: Describing the world in the Reformation*. Aldershot; Burlington, VT: Ashgate, VIII, 378 p. (In English)
- Meijers, D. J. (2005) The paper museum as a genre: The corpus of drawings in St Petersburg within a European perspective. In: R. E. Kistemaker, N. P. Kopaneva, D. J. Meijers, G. V. Vilibakhov (eds.). *The paper museum of the Academy of sciences in St Petersburg c. 1725–1760: Introduction and Interpretation*. Amsterdam: Royal Netherlands Academy of Arts and Sciences, pp. 19–54. (In English)
- Mundt, F. (2008) *Beatus Rhenanus, Rerum Germanicarum libri tres (1531): Ausgabe, Übersetzung, Studien*. Berlin: Walter De Gruyter, 674 p. (In German)
- Ogilvie, B. W. (2006) *The science of describing: Natural history in Renaissance Europe*. Chicago: University of Chicago Press, XVI, 385 p. (In English)
- Sergeev, M. (2019) Der Mithridates (1555) zwischen Sprachmuseum und neulateinischem Onomastikon: einige Überlegungen zur Konzeption und zum Genre des Gessnerschen Handbuchs. In: U. B. Leu, P. Opitz (Hrsg.). *Conrad Gessner (1516–1565): Die Renaissance der Wissenschaften / The Renaissance of Learning*. Oldenburg: De Gruyter, S. 517–532. (In German)
- Sergeev, M. L. (2018) *Sopostavlenie yazykov v XVI veke: na primere “Mitridata” (1555) Konrada Gessnera*. PhD dissertation (Philology). Saint Petersburg, Institute for Linguistic Studies of the Russian Academy of Sciences, 234 p. (In Russian)
- Shumilin, M. V. (2016) “Mitridat” Konrada Gesnera i “Diatriba o yazykakh evropejtsjev” Iosifa Yusta Skaligera: Sravnitel'noe yazykoznanie v XVI veke. In: Ju. V. Ivanova, M. V. Shumilin (eds.). *Nauki o yazyke i tekste v Evrope XIV–XVI vekov*. Moscow: Delo Publ., pp. 169–199. (In Russian)
- Simon, F. (2018) Collecting languages, alphabets and texts: The circulation of “parts of texts” among paper cabinets of linguistic curiosities (sixteenth-seventeenth century). In: F. Bretelle-Establet, S. Schmitt (eds.). *Pieces and parts in scientific texts*. Cham: Springer, pp. 297–346. DOI: 10.1007/978-3-319-78467-0_10 (In English)
- Simon, F. D. (2011) *Sortir de Babel. Une République des langues en quête d'une “langue universelle” à la Renaissance et à l'Âge classique? PhD dissertation (History)*. Rennes, University of Rennes 2, 932 p. (In French)

- Stepanova, L. G. (2000) *Ital'yanskaya lingvisticheskaya mysl' XIV–XVI vekov (Ot Dante do pozdnego Vozrozhdeniya)*. Saint Petersburg: Russian Christian Humanitarian Academy Publ., 498 p. (In Russian)
- Swiggers, P. (1997) *Histoire de la pensée linguistique: Analyse du langage et réflexion linguistique dans la culture occidentale, de l'Antiquité au XIXe siècle*. Paris: Presses Universitaires de France, VIII, 312 p. (In French)
- Tavoni, M. (1986) On the Renaissance idea that Latin derives from Greek. *Annali della Scuola normale superiore di Pisa. Classe di lettere e filosofia. Serie III*, 16 (1): 205–238. (In English)
- Van Hal, T., Considine, J. (2010) Classifying and comparing languages in Post-Renaissance Europe (1600–1800). *Language & History*, 53 (2): 63–69. DOI: 10.1179/175975310X12798962415107 (In English)
- Van Hal, T. (2010) “Moedertalen en taalmoeders”. *Het vroegmoderne taalvergelijkende onderzoek in de Lage Landen*. Brussel: KVAB., IX, 616 p. (In Dutch)
- Van Hal, T., Isebaert, L., Swiggers, P. (2013) Het “vernieuwde” taal- en wererlbeeld van de vroegmoderne tijd. Bakens en referentiepunten. In: T. Van Hal, L. Isebaert, P. Swiggers (eds.). *De tuin der talen: Taalstudie en taalcultuur in de Lage Landen, 1450–1750*. Leuven: Peeters, pp. 3–46. (In Dutch)
- Van Rooy, R. (2017) *Through the vast labyrinth of languages and dialects: The emergence and transformations of a conceptual pair in the early modern period (ca. 1478–1782)*. PhD Dissertation (Linguistics). Leuven, KU Leuven, 571 p. (In English)
- Vorobyev, G. (2018) Sylvia: Zur Entstehung des wissenschaftlichen Namens der Grasmücke (Arist. Hist. an. 592b22). *Philologia Classica*, 13 (2): 247–264. DOI: 10.21638/11701/spbu20.2018.206 (In German)
- Vorobyev, G. M. (2015) Kritika latinskikh perevodov Feodora Gazy v “Istorii zhivotnykh” Konrada Gesnera [Theodore Gaza’s Latin translations criticized in Conrad Gesner’s *Historia animalium*]. *Vestnik Russkoj khristianskoj gumanitarnoj akademii — Review of the Russian Christian Academy for the Humanities*, 16 (2): 333–339. (In Russian)

Сведения об авторе:

Михаил Львович Сергеев, ORCID: [0000-0002-1548-3901](https://orcid.org/0000-0002-1548-3901), e-mail: librorumcustos@gmail.com

Для цитирования: Сергеев, М. Л. (2019) Истории, музеи, словари: формы представления знания о языках в XVI веке и первые лингвистические справочники (книги-полиглоты). *Journal of Applied Linguistics and Lexicography*, 1 (2): 381–389. DOI: 10.33910/2687-0215-2019-1-2-381-389

Получена 17 августа 2019; прошла рецензирование 4 сентября 2019; принята 28 сентября 2019.

Права: © Автор (2019). Опубликовано Российским государственным педагогическим университетом им. А. И. Герцена. Открытый доступ на условиях лицензии CC BY-NC 4.0.

Author:

Mikhail L. Sergeev, ORCID: [0000-0002-1548-3901](https://orcid.org/0000-0002-1548-3901), e-mail: librorumcustos@gmail.com

For citation: Sergeev, M. L. (2019) Histories, museums and dictionaries: the forms of representing linguistic knowledge in the 16th century and the earliest linguistic handbooks. *Journal of Applied Linguistics and Lexicography*, 1 (2): 381–389. DOI: 10.33910/2687-0215-2019-1-2-381-389

Received 17 August 2019; reviewed 4 September 2019; accepted 28 September 2019.

Copyright: © The Author (2019). Published by Herzen State Pedagogical University of Russia. Open access under CC BY-NC License 4.0.

ТОЛКОВЫЙ СЛОВАРЬ-СПРАВОЧНИК КАК УЧЕБНОЕ ПОСОБИЕ

А. А. Хуснутдинов✉¹¹ Ивановский государственный университет, 153025, Россия, г. Иваново, ул. Ермака, д. 39

EXPLANATORY REFERENCE DICTIONARY AS A TEXTBOOK

A. A. Husnutdinov✉¹¹ Ivanovo State University, 39 Ermak Str., Ivanovo, 153025, Russia.

Аннотация. Составление словарей, которые используются в качестве учебных пособий при изучении языка как родного и неродного, представляет собой отдельное направление в лексикографии. Учебные словари имеют свое назначение, своего адресата и особую дидактическую направленность. Этим определяются корпус представляемых в таком словаре языковых единиц, параметры их лексикографического описания, содержание и структура текста словаря и словарных статей, включение в него специальных приложений. От словарей-справочников учебные словари отличает также их избирательность, которая выражается в ориентации только на литературный язык, целенаправленном и строгом отборе круга источников, языковых единиц, параметров их описания и иллюстративного материала. Особенности учебных словарей обусловлена ограниченность сферы их использования и узкий круг пользователей.

Общение современного человека становится все более широким и разнообразным. В процессе коммуникации человек сталкивается с большим количеством неизвестных ему иноязычных слов и выражений, специальных обозначений и терминов из разных областей человеческого знания, именованных, стоящих за пределами литературного языка. Современного пользователя часто интересуют и более подробные сведения о языковых единицах: происхождение и особенности функционирования слова или выражения в языке, возможности использования в разных условиях общения, словообразовательные связи и т. д. Потребность в такого рода информации учебные словари удовлетворить не могут. Поэтому необходимо обратиться к другим типам словарей, в частности к словарям-справочникам. Наблюдения показывают, что толковые словари-справочники обладают определенным образовательным потенциалом и могут использоваться в качестве пособия, дополняющего учебные словари. Среди таких особенностей следует отметить: расширенный словник, позволяющий

Abstract. The compilation of dictionaries to be used as special textbooks when learning a native or a non-native language is a special area of lexicography. Educational dictionaries have their own purpose, audience, and didactic focus. These determine the corpus of the language units described, the parameters of their lexicographic description, the content and structure of the dictionary text and entries, and the inclusion of special appendixes. Educational dictionaries also differ from reference dictionaries in their selectivity, which is expressed in orientation exclusively towards the literary norm, a purposeful and rigorous selection of language sources and language units, along with the parameters of their description and illustrative material. The specificity of educational dictionaries explains the limited scope of their use and the relatively small number of users.

The sphere of modern communication is expanding and becoming more diverse. In the process of communication, a person is confronted with a large number of foreign words and expressions unknown to him, special names and terms from different areas of human knowledge, nominations, which are outside the literary language. A modern user is interested in more detailed information about language units: their origin and peculiar features, the way a word functions or expresses a concept in a language; the possibilities of using it in different communicative spheres, its morphological links, etc.

The educational dictionaries cannot satisfy the overall need for such information. Therefore, it is necessary to refer to other types of dictionaries, in particular to reference dictionaries. Observations suggest that explanatory reference dictionaries have a certain educational potential and can be used as a supplement for educational dictionaries. Among their peculiar features it is necessary to note an extended vocabulary, which provides information on the language units that are not described in educational dictionaries; more detailed information on the form, content and usage of a lexical unit; the specifics of borrowed

получить сведения о языковых единицах, которые не описываются в учебных словарях; более подробную информацию о форме, содержании и употреблении в речи языковой единицы; специфику функционирования заимствованной единицы в родном языке и языке-источнике и др.

Возможность использования толковых словарей-справочников в образовательных целях диктует и необходимость дальнейшего совершенствования современных словарей по ряду характеристик, включающих в себя оперативность обновления информации в словаре, максимальную полноту информации, удобство и доступность представления сведений, вероятность выборочного использования содержащихся в словаре данных и т. д.

Ключевые слова: лексикография, учебная лексикография, толковый словарь, словарь-справочник, учебный словарь, учебное пособие, одноязычный словарь, переводный словарь.

units' functions in one's native language and the source language, etc.

The possible use of explanatory reference dictionaries for educational purposes substantiates the need to further improve contemporary dictionaries in a number of areas, including the update rate, the fullness, availability and convenience and information, selective use of the data contained in the dictionary, etc.

Keywords: lexicography, educational lexicography, explanatory dictionary, reference dictionary, educational dictionary, manual, monolingual dictionary, translated dictionary.

Учебные словари и потребности современного пользователя

Словарь как учебное пособие для овладения языком — как родным, так и неродным — используется не одно столетие. В наши дни учебная лексикография составляет особое направление в словарном деле, отличающееся от других своими целями и задачами, своим адресатом, своим кругом теоретических и практических проблем, которыми определяются, в конечном счете, и само своеобразие, и особенности содержания и структуры словарей такого типа. Специфика учебных словарей заключается в их дидактической направленности, а именно в том, что они нацелены на формирование и обогащение активного лексикона пользователя. Поэтому в реестр (словник) учебного словаря отбираются такие языковые единицы, которые находятся в активном употреблении в современном литературном языке, а включенные в словарь единицы описываются с тех сторон и с той степенью полноты и детальности, которая позволяет пользователю не только получить представление об особенностях формы и содержания той или иной единицы, но и активно использовать ее в своей собственной речи в различных условиях общения и в соответствии с принятыми нормами и правилами. Этим учебные словари отличаются от словарей-справочников, которые предназначены для предоставления сведений о тех или иных сторонах описываемых в словаре языковых единиц. Дидактической направленностью определяется своеобразие самого учебного словаря и его место среди других типов лексикографических изданий. Избирательность учебных словарей, выражающаяся в ориентации только на литературный язык, целенаправленном и строгом отборе источников, самих языковых единиц, параметров их описания, а также в объеме и способах подачи материала, обуславливает ограниченность сферы использования учебных словарей и весьма узкий круг пользователей.

Общение современного человека становится все более широким и разнообразным. Если раньше эта сфера определялась в значительной степени социальным положением, профессиональной деятельностью и была, как правило, довольно узкой, то сейчас каждый имеет практически неограниченные возможности для удовлетворения своих потребностей в любой информации и обмене ею. Вследствие этого наш современник в процессе общения сталкивается с самыми разными по форме, содержанию и целевому назначению речевыми произведениями (текстами). Это не только художественные тексты разных типов и жанров,

публицистика, но и научные труды, официально-деловые документы и т. д. Общение в сети Интернет отражает и живую устную речь нашего времени во всем многообразии. При этом современный человек сталкивается с большим количеством незнакомых ему иноязычных слов и выражений, специальных обозначений и терминов из разных областей человеческого знания, именованных, стоящих за пределами литературного языка, а также общеупотребительных слов, использующихся в значениях, которые ему неизвестны. Так, анализ употребления слова **железо** в русском национальном языке демонстрирует весьма широкий диапазон значений: **железо** — это ‘химический элемент’, ‘лекарственный препарат’, ‘металл’, ‘металлическое изделие’, ‘холодное оружие (обычно меч или нож)’, ‘огнестрельное оружие (пистолет, автомат)’, ‘орудия охоты (капканы)’, ‘орудия пыток’, ‘кандалы’, ‘металлические орудия труда и инструменты’, ‘спортивное снаряжение, снаряды и т. п.’, ‘культуризм’, ‘музыкальные инструменты (тарелки)’, ‘динамики’, ‘металлический рок’, ‘металлические зубы’, ‘монеты’, ‘деньги (валюта)’, ‘металлические части оружия, орудий и т. п.’, ‘металлические части электрогитары’, ‘механизмы из железа’, ‘автомобиль’, ‘старый, разбитый, не годный для эксплуатации автомобиль’, ‘корабль’, ‘флот’, ‘аппаратная часть компьютера’, ‘надежный человек’, ‘суровый, непреклонный человек’, ‘холодный, бесчувственный человек’.¹ Как видим, слово **железо** в современной речи используется не только в общеупотребительных значениях, которые обычно фиксируются словарями литературного языка, но и в таких, в которых оно имеет ограниченное употребление за пределами литературного языка (в просторечии: ‘металлические зубы’, ‘автомобиль’ и др.; в сленговой речи: ‘спортивное снаряжение, снаряды и т. п.’, ‘культуризм’ (спорт.), ‘музыкальные инструменты’, ‘металлический рок’ (муз.), ‘корабль’, ‘флот’ (флотск.) и т. д.). Понимание текстов, в которых общеупотребительные слова используются в таких значениях, невозможно без обращения к толковым словарям. Обращения к словарям требуют и частотные в современной живой речи иноязычные слова и выражения. Приведем в качестве примера некоторые высказывания, выбранные из сети Интернет, в которых используется слово **фейспалм (facepalm)**: *Единственное чувство во время просмотра фильма — бесконечный **фейспалм** от той чуши, которая на экране происходит. Я едва удержался от того, чтобы не изобразить **фейспалм**. Начало у фильма бурное и энергичное, но уже в нем я начал ловить **facepalm**, и не один. **Facepalm** или когда же Путин распустит Думу? Настоящий **фейспалм**: iPhone X не узнал лицо своего хозяина на презентации* и др. В таких случаях человеку для понимания чужих текстов и полноценного общения необходимы хотя бы самые общие сведения о форме и значении многих используемых в речи слов и выражений.

Современного пользователя часто интересуют и более подробные сведения о языковых единицах: происхождение и особенности функционирования слова или выражения в языке, возможности использования в разных условиях общения, парадигматические связи с другими словами, словообразовательные потенции и т. д. Очевидно, что потребность в информации такого рода учебные словари, в силу их специфики, удовлетворить, как правило, не могут. Поэтому оказывается необходимым обращение к другим типам словарей, в частности к словарям-справочникам.

Толковые словари-справочники как особый тип филологического словаря

Толковые словари следует, вероятно, отнести к самому распространенному типу лексикографических изданий. Они не только многочисленны, но и разнообразны. Толковые словари могут быть нормативными (словари литературного языка) и ненормативными

¹ Подробнее о значении и употреблении слова **железо** в русском языке см. (Хуснутдинов, Хуснутдинова 2012).

(описательными, фиксирующими реальное употребление), большими, средними и малыми по объему (многотомными, в несколько томов и одготомными), словарями-справочниками и учебными, синхронными и диахронными (историческими), одноязычными и неоднотязычными (переводными) и т. д. Основное назначение толковых словарей — предоставление справок (сведений) о языковой единице в целом или о ее отдельных свойствах. Каждое свойство (признак) языковой единицы разрабатывается в словаре как отдельный лексикографический параметр. Пользователям словаря обычно необходима информация о разных сторонах языковой единицы. Поэтому особенностью толковых словарей является их многопараметровость — описание включенных в словник словаря языковых единиц по целому ряду параметров, в первую очередь таких, которые относятся к их форме, значению и употреблению в речи. В число таких параметров входят, как правило, написание слова, произношение и ударение, толкование значения, частеречная принадлежность и грамматические свойства, лексическая и грамматическая сочетаемость со словами в речи, сфера бытования, стилистическая и эмоционально-экспрессивная окрашенность, историко-временная отнесенность, происхождение и др.

Полнота описания, выражающаяся в широте охвата описываемого языкового материала, многосторонности и глубине лексикографической разработки включенных в словарь единиц, определяет востребованность толкового словаря у пользователя и степень его удовлетворенности качеством словаря. Для современного пользователя особенно ценными, а часто и незаменимыми пособиями при чтении художественных и нехудожественных текстов XIX–XX веков (т. е. таких текстов, время создания которых отдалено от современного читателя значительным промежутком) оказываются такие словари, как «Толковый словарь живого великорусского языка» В. И. Даля, «Толковый словарь русского языка» под ред. Д. Н. Ушакова (ТСУ), «Словарь современного русского литературного языка» в 17 томах (БАС), издающийся сейчас «Большой академический словарь русского языка». Следует особо указать, что научная и практическая ценность таких словарей, отражающих язык своей эпохи, со временем только возрастает.

В толковых словарях, и не только толковых учебных, заключен большой образовательный потенциал, так как толковый словарь более всех других типов словарей может способствовать достижению таких образовательных целей, как увеличение активного и пассивного словарного запаса человека, расширение его кругозора и углубление его знаний, в том числе и лингвистических, возрастание интереса к языку, в первую очередь к родному, уделение большего внимания качеству своей и чужой речи. Образовательный потенциал толковых словарей обычно хорошо осознается составителями. Так, В. И. Даль прямо указывал на то, что словарь должен служить, помимо всего прочего, пособием для совершенного овладения родным языком, ср. его суждения на этот счет:

«Словарь толковый <...> собственно и должен служить для изучения родного языка, для отыскания всех выражений, какие могут кому понадобиться, для сравнения тождесловов и пр.; словом, я себе воображаю словарь этот настольною книгою каждого образованного человека» (Даль 1989, т. 1, ХС);

«Толковый словарь, расположенный по предметам, так что легко было бы отыскать каждое потребное слово, каждое выражение, узнать сравнительное значение его, замену его тут и там другим, видеть в то же время пред собою путные примеры оборотов речи, правильного русского склада, окинуть в небольшой статье весь запас речений нашего богатого языка, для выражения данного понятия — это потребность насущная <...>» (Даль 1989, т. 1, ХС1).

Направленность толкового словаря на обучение особенно заметна в «Толковом словаре русского языка» под ред. Д. Н. Ушакова, который создавался с установкой «для поль-

зования и учения всех»². Эта образовательная и нормализаторская направленность прямо выражена в разделе «Как пользоваться словарем» (ТСУ, т. I, XXII–LXXVI). В нем не только сообщается о содержании и построении словаря, но и даются сведения по правописанию, произношению и грамматике. Этим же целям служат списки слов, требующих особых комментариев (числительные, разряды прилагательных и наречий). К словарю также прилагается список иноязычных слов и выражений, которые в русском языке употребляются без перевода (ТСУ, т. IV, 1472–1484).

Образовательный потенциал имеется в каждом толковом словаре, поэтому словари-справочники этого типа должны широко использоваться в практике обучения языку (и родному, и неродному). Важно также сформировать в процессе обучения навыки пользования словарем, а именно умение выбрать нужный в данный момент словарь и извлечь из него необходимую информацию. Для этого и обучающему, и обучаемому следует ясно представлять себе образовательный потенциал толкового словаря-справочника как особого типа информационно-справочного издания и каждого конкретного словаря, имеющегося в обиходе.

Образовательный потенциал толкового словаря

Преимущество толкового словаря-справочника заключается прежде всего в расширенном по сравнению со словарем учебным словнике: количество описываемых в словаре-справочнике слов, как минимум, на порядок больше, чем в словаре учебном. Отсюда более широкий тематический охват лексики и, соответственно, более полное отражение понятийной сети языка. Словарь-справочник дает пользователю возможность представить количественное и качественное многообразие языковых единиц, входящих в словарный состав языка, разнообразие связей и отношений, в которые вступают эти единицы друг с другом в системе языка. Важным здесь является и удовлетворение практической потребности найти нужное слово в словаре и получить необходимые сведения о нем. Особенно это касается языковых единиц, имеющих какие-либо ограничения в употреблении, связанные с отнесенностью их к определенной тематической группе, просторечию, сленгу и т. д.

Специфика словарей-справочников, а именно многопараметровость описания, определяет и их другую — важную для нашей темы — особенность, а именно возможность дать более полную и разностороннюю характеристику включенного в реестр слова. Это касается в первую очередь отражения в словаре вариантных форм, в которых слово реально употребляется в речи, значений и оттенков значений, которые по тем или иным причинам не включаются в лексикографическое описание лексической единицы в учебных словарях, а также таких особенностей, которыми определяются возможности использования данного слова в разных речевых ситуациях и контекстах. Существенными в этом отношении представляются такие параметры описания единицы, как сочетаемость с другими словами в речи (лексическая и грамматическая), сфера употребления (литературный язык, просторечие, территориальные и социальные диалекты), стилистическая принадлежность (разговорное, книжное), эмоционально-экспрессивная окраска (одобрительно, пренебрежительно, шуточно и т. п.), историко-временная отнесенность (актуальное, устаревшее, новое), а также эпидигматические связи, которые отражают «способность слова, благодаря словообразованию и процессам его семантического развития, входить одновременно в различные лексико-семантические парадигмы» (Ярцева 1990, 367). Такая информация важна для пользователя, так как она позволяет ему представить диапазон возможного использования слова в речи и сделать выбор в соответствии с конкретными условиями общения.

² Об истории создания и особенностях этого словаря см., например: (Сороколетов 1998, 346–366).

В этой связи следует указать и на достаточно обширный и разнообразный иллюстративный материал толковых словарей-справочников, в котором представлены примеры правильного и образцового употребления слова в речи.

Установка толковых словарей-справочников на полноту охвата лексики и всестороннее описание единицы существенно снижает избирательность в отборе и описании языковых единиц, которая характерна для учебных словарей. Отказ от такой избирательности дает возможность не просто представить словарный состав языка в возможной его полноте, но и показать его в системном виде, т. е. не только описать каждое слово в отдельности, но и указать на его связи и отношения с другими единицами (семантические, грамматические, деривационные, синтагматические и др.). Эта особенность толковых словарей-справочников способствует расширению и углублению знаний пользователя о языке и языковых единицах, что, в свою очередь, помогает ему осознанно выбирать языковые средства и формы выражения в различных условиях общения, контролировать и регулировать свое речевое поведение.

Представляется важным и то, что толковый словарь-справочник способен показать динамические процессы, происходящие в словарном составе языка в целом и каждой единице в отдельности. Толковый словарь-справочник может фиксировать и комментировать изменения, которые претерпевают слова в процессе функционирования в языке: какие из них остаются неизменными на протяжении длительного времени, с какими происходят изменения, касающиеся их формы, значения и других свойств. Знание таких сведений важно для грамотного употребления языковых единиц в речи.

Особое значение одноязычные толковые словари-справочники приобретают при изучении иностранного языка и переводе текстов с одного языка на другой. Переводные словари, в том числе и учебные, нацелены на установление соответствий слов в разных языках, иначе говоря, на установление общего, сходного, эквивалентного. При установлении таких соответствий в переводных словарях, как иноязычно-русских, так и русско-иноязычных, нивелируются индивидуальные особенности соотносимых слов в каждом языке, так как акцент в них делается именно на установление общих черт. В одноязычных словарях, наоборот, основной задачей является как описание формальных и содержательных свойств слова, так и показ особенностей функционирования слова в каждом языке.³ Поэтому использование одноязычных словарей-справочников при переводе и изучении другого языка имеет не только чисто практическое значение: оно формирует общую филологическую эрудицию и лингвистическую компетентность.

Уже представленный в этом разделе обзор особенностей словарей-справочников показывает их значительный образовательный потенциал, которым определяется возможность и необходимость использования их в практике обучения и самообразования. Здесь встает во весь рост особая задача, которую можно определить как формирование лексикографической грамотности пользователя, выработки у него умения, навыков и привычки пользоваться словарями. Эта тема весьма обширна и требует отдельного обсуждения.

Здесь же следует сказать еще об одной задаче, а именно о необходимости специального изучения образовательного потенциала неучебных словарей разных типов (и в первую очередь толковых словарей-справочников). Такое изучение должно быть ориентировано, на наш взгляд, не только на решение задачи формирования лексикографической грамотности пользователей, но и на оценку образовательного потенциала существующих словарей

³ Отсутствие полной эквивалентности в реальном употреблении значительного количества слов, которые определяются в переводных словарях как эквиваленты, легко устанавливается при соотнесении словарных статей на данное слово в одноязычных словарях. Специально на фразеологическом материале эквивалентность русских и немецких идиом в словаре и тексте исследована К. Шиндлер (Schindler 2005).

справочного типа и на учет образовательных потребностей современных пользователей при разработке новых лексикографических проектов. Укажем на некоторые из них.

Какой толковый словарь нужен современному пользователю

Следует прежде всего указать на то, что «бумажные» словари, очень удобные для учебных целей, по ряду причин не могут удовлетворить современного пользователя. В особенности это касается словаря толкового, который в бумажном варианте все более отстает от потребностей современного пользователя. В первую очередь это касается оперативности словаря: каждое новое слово или выражение, а также изменения, происходящие с ним в живой речи, должны в максимально короткий срок фиксироваться и отражаться в словаре. Такая оперативность может быть достигнута только в электронных онлайн-словарях. Следовательно, современный толковый словарь требует не только разработки и реализации проекта, но и постоянного его сопровождения, предусматривающего оперативное внесение в словарь необходимых дополнений и поправок, что позволило бы пользователю иметь текст словаря в последней и новейшей редакции.

Ценным для пользователя качеством словаря является удобство доступа к нему. Это касается не только «технической» стороны, например доступности в телефоне, но и собственно содержания и построения текста словаря, особенностей его электронной «оболочки», способов подачи информации, метаязыка словаря и т. д. Здесь важно также, чтобы словарь сохранял в содержании научную объективность и точность, но по форме изложения был простым и ясным, доступным широкому кругу пользователей с разным уровнем образования и лингвистической подготовки.

Важным свойством любого толкового словаря остается его полнота — полнота «внешняя», выражающаяся в стремлении охватить по возможности весь корпус единиц, описание которых ставит задачей словарь, и полнота «внутренняя», которая обеспечивает многопараметровость описания включенных в словник словаря единиц. Расширение круга пользователей и разнообразие необходимой им информации приводят не только к расширению словника словаря, но и к увеличению количества параметров описания единицы. Так, современному пользователю часто недостаточно получить сведения о значении слова, ему интересна и информация энциклопедического характера о самом объекте, обозначенном словом. В отдельных случаях не лишними для правильного употребления слова в речи оказываются сведения о его происхождении и функционировании в языке. Например, использование в русском языке слова **магометанин** для обозначения лица, исповедующего ислам, до XVII века было оправданным в связи с отсутствием непосредственных контактов с мусульманскими народами; употребление этого слова в современной речи является недопустимым, так как для мусульман оно звучит оскорбительно, потому что указывает на поклонение мусульман не Аллаху, а пророку Магомету, т. е. человеку, что в корне противоречит сущности исламской религии (ср. у христиан: **Христос, христианин**).

Существенным для оценки полноты толкового словаря является включение в лексикографическое описание слова употребление его в составе устойчивых сочетаний различных типов: идиом, обозначений терминологического характера, пословично-поговорочных выражений и т. д. Включение в словарь таких сочетаний и их описание не только демонстрирует диапазон использования слова в языке, но и обогащает лексикон пользователя определенным количеством устойчивых выражений.

Современный пользователь оценит в толковом словаре и рекомендации нормативного характера. Для пользователя с широким и разнообразным кругом общения оказывается недостаточным простое указание на соответствие или несоответствие современным нормам литературного языка. Для него важны сведения, указывающие на возможность/невозмож-

ность, допустимость/недопустимость, уместность/неуместность того или иного употребления в разных условиях общения и конкретных речевых ситуациях.

Особую значимость для пользователя содержащиеся в толковом словаре сведения приобретают при необходимости перевода слова: объем информации о языковых единицах в одноязычном толковом словаре значительно больше, чем в переводных словарях. Одноязычные толковые словари дают возможность соотнести слово не только в целом, но и по каждому из параметров в отдельности, что позволяет достаточно точно установить степень эквивалентности слов разных языков.

Выводы

Образовательный потенциал, заключенный в толковом словаре-справочнике, всегда осознавался составителями, однако он по большей части оказывался на втором плане, потому что применение толкового словаря в образовательных и учебных целях оставлялось на усмотрение пользователей (составители словарей чаще всего ограничивались указанием на нормативность/ненормативность того или иного употребления).

Использование толковых словарей-справочников в образовательных и учебных целях должно стать предметом специального обсуждения не только в преподавательской среде, но и в кругу лексикографов — теоретиков и практиков словарного дела. Иначе говоря, образовательный потенциал целенаправленно должен закладываться в толковый словарь-справочник уже на стадии его разработки. Необходима также разработка методики работы со словарем в учебных и образовательных целях для различных категорий пользователей (школьники, студенты, иностранцы и др.).

Специальной задачей является работа, направленная на повышение общей лексикографической грамотности пользователей, на выработку умения и привычки обращаться к словарям в обиходе и профессиональной деятельности.

Словари

Даль, В. И. (1989–1991) *Толковый словарь живого великорусского языка*. В 4 т. М.: Русский язык.
 ТСУ — Ушаков, Д. Н. (ред.). (1935–1940) *Толковый словарь русского языка*. В 4 т. М.: Советская энциклопедия.
 Ярцева, В. Н. (ред.). (1990) *Лингвистический энциклопедический словарь*. М.: Советская энциклопедия, 685 с.

Литература

Даль, В. И. (1989) О наречиях русского языка. В кн.: В. И. Даль. *Толковый словарь живого великорусского языка*. В 4 т. Т. 1. М.: Русский язык, с. XLIX–XCIII.
 Сороколетов, Ф. П. (ред.). (1998) *История русской лексикографии*. СПб.: Наука, 610 с.
 Хуснутдинов, А. А., Хуснутдинова А. А. (2012) Лексикографический портрет слова железо. *Вестник Ивановского государственного университета. Серия: Гуманитарные науки*, 1 (12): 54–74.
 Schindler, Ch. (2005) *Untersuchungen zur Äquivalenz von Idiomen in Sprachsystem und Kontext: Am Beispiel des Russischen und des Deutschen*. Münster: LIT Verlag, 275 S. (Veröffentlichungen des Slavisch-Baltischen Seminars der Universität Münster, Bd. 9)

Dictionaries

Dal', V. I. (1989–1991) *Tolkovyy slovar' zhivogo velikorusskogo yazyka*: In 4 vols. Moscow: Russkij yazyk Publ. (In Russian)
 Ushakov, D. N. (ed.). (1935–1940) *Tolkovyy slovar' russkogo yazyka*: In 4 vols. Moscow: Sovetskaya entsiklopediya Publ. (In Russian)

Yartseva, V. N. (ed.). (1990) *Lingvisticheskiy entsiklopedicheskiy slovar'*. Moscow: Sovetskaya entsiklopediya Publ., 685 p. (In Russian)

References

- Dal', V. I. (1989) O narechiyakh russkogo yazyka. In: V. I. Dal'. *Tolkovyy slovar' zhivogo velikorusskogo yazyka: In 4 vols. Vol. 1.* Moscow: Russkiy yazyk Publ., pp. XLIX–XCIII. (In Russian)
- Khusnutdinov, A. A., Khusnutdinova, A. A. (2012) Leksikograficheskiy portret slova *zhelezo* [Lexicographic portrait of a word “iron”]. *Vestnik Ivanovskogo gosudarstvennogo universiteta. Seriya: Gumanitarnye nauki — Ivanovo State University Bulletin, Series “The Humanities”*, 1 (12): 54–74. (In Russian)
- Schindler, Ch. (2005) *Untersuchungen zur Äquivalenz von Idiomen in Sprachsystem und Kontext: Am Beispiel des Russischen und des Deutschen*. Münster: LIT Verlag, 275 S. (Veröffentlichungen des Slavisch-Baltischen Seminars der Universität Münster, Bd. 9) (In German)
- Sorokoletov, F. P. (ed.). (1998) *Istoriya russkoj leksikografii*. Saint Petersburg: Nauka Publ., 610 p. (In Russian)

Сведения об авторе:

Арсен Александрович Хуснутдинов, e-mail: arsen1418@mail.ru

Для цитирования: Хуснутдинов, А. А. (2019) Толковый словарь-справочник как учебное пособие. *Journal of Applied Linguistics and Lexicography*, 1 (2): 390–398. DOI: 10.33910/2687-0215-2019-1-2-390-398

Получена 14 мая 2019; прошла рецензирование 25 августа 2019; принята 28 августа 2019.

Права: © Автор (2019). Опубликовано Российским государственным педагогическим университетом им. А. И. Герцена. Открытый доступ на условиях лицензии CC BY-NC 4.0.

Author:

Arsen A. Husnutdinov, e-mail: arsen1418@mail.ru

For citation: Husnutdinov, A. A. (2019) Explanatory reference dictionary as a textbook. *Journal of applied linguistics and lexicography*, 1 (2): 390-398. DOI: 10.33910/2687-0215-2019-1-2-390-398

Received 14 May 2019; reviewed 25 August 2019; accepted 28 August 2019.

Copyright: © The Author (2019). Published by Herzen State Pedagogical University of Russia. Open access under CC BY-NC License 4.0.

УДК 81`32

НЕЙРОННЫЕ СЕТИ И КОМПЬЮТЕРНАЯ ОБРАБОТКА ЯЗЫКА

О. В. Митренина^{✉1}

¹ Санкт-Петербургский государственный университет, 199034, Россия,
Санкт-Петербург, Университетская наб., д. 7/9

ARTIFICIAL NEURAL NETWORKS AND NATURAL LANGUAGE PROCESSING

O. V. Mitrenina^{✉1}

¹ St Petersburg State University, 7/9 Universitetskaya Emb., Saint Petersburg 199034, Russia

Почему дети учат язык лучше, чем взрослые

Однажды несколько монахов, живших на краю Скитской пустыни, обнаружили у себя корзину. В корзине плакал чернокожий младенец, который, несомненно, был подкинут эфиопским караваном, проходившим тут накануне. Растроганные таким непредвиденным подарком небес, братья стали кормить младенца и усердно заботиться о нем.

Шло время. И вот как-то один из монахов, весьма обеспокоенный, сказал:

— Нужно, чтобы кто-нибудь из нас выучил эфиопский язык.

— Но почему? — воскликнули изумленные братья.

— Потому что скоро младенцу исполнится год, и он начнет говорить, а никто из нас не знает его языка.

В этой истории из книги «Отцы-пустынники смеются» монах предполагал, что ребенок хранит в голове язык своих родителей. На самом деле, дети усваивают тот язык, который слышат вокруг себя. Взрослый человек теряет детскую способность усваивать и вынужден зубрить. Он может бесконечно повторять грамматику и лексику, но так и не научится говорить на новом языке как на родном.

Почему же маленькие дети учат язык таким способом, который уже недоступен взрослым?

Может быть, дело в том, что ребенок усваивает свой первый язык на «чистую голову», а голова взрослого уже заполнена родным языком, и новый язык помещается туда очень плохо? Нет, это не главная причина. Ребенок может расти в двуязычной среде, и тогда он станет билингом — легко усвоит два языка как родные, а может усвоить и три. Без всякой зубрежки. Достаточно, чтобы с ним и вокруг него много говорили на этих языках — они не перепутаются и спокойно «поместятся в голове».

Но с годами способность усваивать язык исчезает даже тогда, когда ни один язык не выучен. Известны истории, когда дети росли вне языковой среды. Девочку под кодовым именем Изабелла обнаружили в шестилетнем возрасте. До шести лет она не слышала человеческой речи. Через год она уже полностью овладела языком и пошла в обычную школу. Другую девочку, Джини, нашли, когда ей было тринадцать. Несмотря на титанические усилия специалистов, она так и не научилась строить предложения более чем из двух или трех слов. Языковые навыки у нее так и не выработались.

Разгадка связана с разницей в устройстве мозга взрослого человека и ребенка. Мозг состоит из нейронов. Это нервные клетки, которые способны передавать информацию друг другу. Для этого между нейронами должны установиться связи. У новорожденного младенца много нейронов, но мало связей между ними. Эти связи с огромной скоростью начинают возникать по мере того, как ребенок знакомится с внешним миром и осваивает новые навыки. В мозгу у него с огромной скоростью выстраиваются нервные цепочки — нейронные пути, нейронные сети. Затем эти нейроны и созданные ими сети начинают покрываться особым веществом, которое называется миелин. Так закрепляется нейронная сеть. Она становится более прочной и удобной, но менее гибкой. Знания фиксируются. Мы начинаем хуже учиться, зато более эффективно использовать накопленный жизненный опыт.

К семи годам выработка миелина уменьшается, сети закрепляются хуже, но к этому времени наш организм успевает выработать основные жизненные навыки. Не все, конечно. Но «глубинные» — почти все. Далее можно переучиваться и доучиваться, но на это будут требоваться дополнительные усилия.

Однако не все потеряно: новые связи продолжают образовываться. Особенно если мы их приучаем образовываться — заставляем себя пробовать что-то новое и учиться. Даже поход домой менее привычной дорогой оживляет процесс создания новых нейронных цепочек. Опасно ограничивать себя только привычными ситуациями и избегать тех, для которых требуются внутренние усилия.

Первые попытки научить компьютер человеческому языку

Если в научно-фантастическом фильме появляется робот, он обязательно говорит на человеческом языке. И это не удивительно — если в фильме он думает и ходит, почему бы ему еще и не разговаривать. Ведь это наиболее простой и естественный механизм коммуникации с человеком.

Но в жизни все оказалось сложнее. Роботы пока не научились свободно общаться с людьми. Глобальная задача «научить компьютер человеческому языку» заменилась на более скромные пожелания из серии «пусть поможет хотя бы в этой области». Вот только некоторые из них:

- переводить тексты с одного языка на другой, чтобы каждый раз не нанимать переводчика;
- просмотреть тысячу-другую блогов и составить отчет о том, что говорят о вашем новом продукте, чтобы не нанимать трех застревающих в сети менеджеров;
- то же самое, но изучая не содержание сообщений, а настроение пишущих — в конце концов, содержание забывается, а настроения сохраняются надолго;
- отвечать на вопросы клиентов — вопросы у них повторяются, и один чат-бот может заменить два десятка менеджеров по работе с клиентами.

Вначале машины пытались научить теми же методами, какими обучают взрослых людей. В их память загружали слова и правила соединения этих слов. Если глагол ПРОСНУТЬСЯ соединить с существительным КОШКА и поставить в прошедшее время, то получится КОШКА ПРОСНУ-ЛАСЬ. А если кошек несколько? Тут форма слова зависит от числа. ТРИ КОШК-И, ПЯТЬ КОШ-ЕК, ДВАДЦАТЬ ОДНА КОШ-КА. Побольше правил, и компьютер научится выдавать человеческие предложения. Будет говорить, как взрослый человек, который учил новый язык по книжкам и не жил в той среде, где на этом языке общаются.

Так работает **подход, основанный на правилах**. Компьютеру дают набор слов и правила их обработки. Это его «знания». Накопив их, он получает предложения, которые должен обрабатывать. Он находит в своем словаре нужные слова или фрагменты слов и применяет к ним правила из своего списка правил.

В итоге компьютер осваивает язык для какой-то конкретной задачи, но совершает много ошибок. Прикладной лингвист, словно опытный учитель, анализирует ошибки своего ученика, дает ему дополнительные правила и новые слова.

При таком способе обучения не используется главное преимущество компьютера — умение хранить и быстро обрабатывать огромные массивы информации (чисел). Человеку трудно сложить в уме восемь десятизначных чисел, а мой простенький ноутбук за одну секунду успеет сделать это сто миллионов раз.

Кто учит лазать по деревьям птицу, которая умеет летать? Так и компьютеру в конце концов предложили другое обучение, не человеческое, а машинное. Так появился **статистический подход**. Это были еще не нейросети, но это был уже «компьютерный» способ анализа языка, а не «человеческий».

Язык не так случаен, как кажется

Но если не использовать правила грамматики, то что может узнать компьютер о языке?

Вспомним Шерлока Холмса. В рассказе «Пляшущие человечки» он разгадал надписи, которые все принимали за детские рисунки. Вот первый текст, который попал ему в руки:

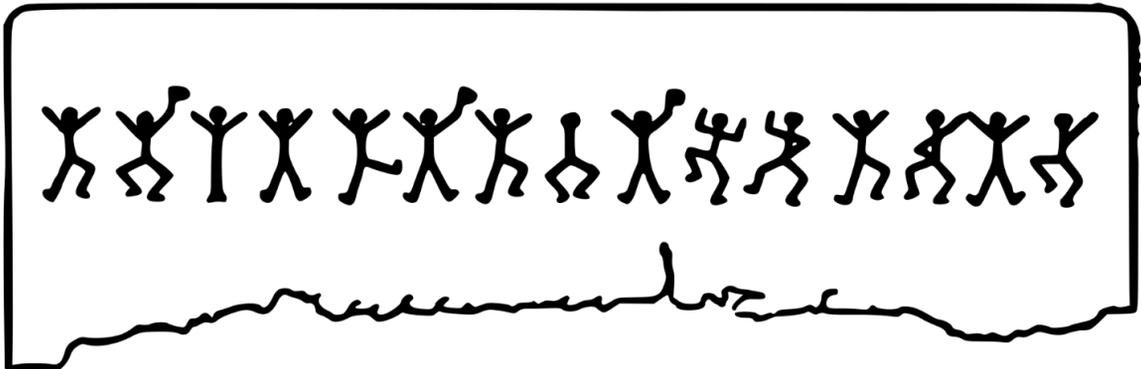


Рис. 1. Первое зашифрованное письмо, оказавшееся у Шерлока Холмса

(Conan Doyle, A. (1905) The adventure of the dancing men. In: *The return of Sherlock Holmes*. New York: McClure, Phillips & Co, pp. 61–92. URL: <https://archive.org/details/returnofsherlock00doyliala/page/64/mode/2up/>)

Холмс догадался, что флажки разбивают цепочку на слова. Он вычислил, какой человечек встречается чаще всего, и предположил, что это буква Е. Она в английском языке самая частая:



Рис. 2. Самый частый человечек из первого письма

(Conan Doyle, A. (1905) The adventure of the dancing men. In: *The return of Sherlock Holmes*. New York: McClure, Phillips & Co, pp. 61–92. URL: <https://archive.org/details/returnofsherlock00doyliala/page/64/mode/2up/>)

После этого Холмс попытался понять, в каком окружении встречается Е, и это подход компьютерного лингвиста. Но имевшийся у него набор текстов был слишком мал — всего одна надпись. Лишь получив еще две надписи, Холмс нашел слово из пяти букв, которое начиналось и заканчивалось буквой Е. Он вспомнил имя хозяйки — *ELSIE* — и вычислил еще три буквы, а потом разгадал и все остальное.

Тут важно, что Холмс не использовал ни правила, ни грамматику языка, а оперировал только статистическими данными о языке.

Компьютер, как и Холмс, может находить закономерности в текстах. Для этого он должен усвоить некоторые знания о языке. Ведь Холмс заранее знал, что самая частая буква — Е и что слово *ELSIE* вполне может встретиться в записке.

Что и как может узнать компьютер?

Откуда компьютер может узнать о языке, если не давать ему правил из школьной грамматики? Точно так же, как и ребенок, компьютер может узнать о свойствах языка прямо из текстов. Для этого ему требуется много текстов, признаки для анализа и алгоритм обучения, по которому он может что-то узнать. Рассмотрим каждый из этих трех пунктов внимательнее.

(1) Много текстов, на которых будет учиться компьютер. Набор таких текстов называется **обучающим корпусом**. Корпус может быть размеченным или нет. В размеченном корпусе заранее указаны «готовые ответы», а иногда и признаки текстов. Например, обучающим корпусом может стать большая коллекция электронных писем, каждое из которых имеет пометку «не спам» или «спам». Эти пометки и являются «готовыми ответами»; изучив их, компьютер научится вылавливать новые письма со спамом.

Обучающим корпусом с «готовыми ответами» может стать большой набор диалогов. Его можно рассматривать как пары «реплика — готовый ответ на реплику». Обработав их, компьютер научится поддерживать беседу. Как и человек, оказавшись в новом для него обществе — среди разбойников, например, или среди врачей — может послушать их язык, а затем научиться «говорить как они».

Если мы хотим заставить компьютер переводить с одного языка на другой, то надо дать ему большой обучающий корпус переведенных текстов: много предложений на исходном языке и их переводы. И никакой грамматики, заметьте! В качестве «готовых ответов» здесь выступают готовые переводы. Иногда в такой корпус добавляли признаки: устанавливали соответствия между словами или словосочетаниями. И напрасно. Как оказалось, без них машина переводит лучше.

Иногда для обучения используется неразмеченный корпус — просто тексты без всяких отметок. Тогда машина должна сама находить в них признаки и закономерности. Именно с нейросетями она научилась делать это весьма хорошо.

(2) Признаки для анализа. Но если компьютер должен обработать множество текстов, то что именно он должен анализировать, какие параметры?

Холмс, расшифровывая пляшущих человечков, использовал признак частоты букв, а также некоторый хранившийся у него в голове словарь. Компьютеру тоже нужны **признаки** — особенности, характерные черты, — с помощью которых он будет анализировать тексты. Например, для оценки тональности текста пригодится словарь: если в сообщении есть слово ГАДОСТЬ, то, вероятно, это сообщение с сильной отрицательной тональностью. А если слово НЕПЛОХОЙ, то тональность, скорее всего, умеренно-положительная. Но признаки могут быть и менее очевидными. Например, при распознавании лиц компьютер будет анализировать признаки изгиба линий — такие, которые описать словами было бы

невозможно. В текстах, как и в лицах, есть свои неявные признаки, которые компьютер сможет выявить.

(3) Алгоритм обучения. Как компьютер должен обрабатывать язык на основе обучающего корпуса и признаков? Так же, как и мозг ребенка: он должен находить не «правильный ответ» на основании четких правил, а «наиболее вероятный» ответ на основании уже полученных примеров правильных конструкций. В большинстве случаев этого бывает достаточно. Компьютерные алгоритмы постоянно совершенствуются, и все более повышается вероятность правильного ответа.

Рассмотрим на примере, как компьютер использует вероятностные алгоритмы, чтобы научиться обращаться с языком.

В известном анекдоте человек после вечеринки садится в такси и на вопрос таксиста «Куда вам?» отвечает:

— К удавам не хочу.

— КУДА ВАМ НАДО?! — четко выговаривая слова, спрашивает таксист.

— Ну, надо так надо... Поехали к удавам, — обреченно соглашается пассажир.

Наш жизненный опыт подсказывает, что фраза «Куда вам надо» гораздо более вероятна, чем «К удавам надо». Особенно в контексте диалога с таксистом. Компьютер тоже может выбирать из нескольких возможных предложений наиболее вероятное. Это полезно, например, при составлении субтитров к видеозаписи или при автоматическом переводе. Можно не объяснять машине, что в предложении КОШКА ПРОСНУЛСЯ сказуемое неправильно согласуется с подлежащим по роду. Проще использовать вероятность. Сочетание КОШКА ПРОСНУЛАСЬ более вероятно, чем КОШКА ПРОСНУЛСЯ.

Как оценить вероятность предложения на основе корпуса? Самый простой способ — просто посмотреть, какие предложения встречались чаще. Но это плохой способ. Даже если корпус очень большой, он не охватывает всех возможных предложений.

Другой способ — использовать вероятности входящих в предложение слов. Какие-то слова встречаются чаще, какие-то — реже. С помощью корпуса можно легко посчитать вероятности слов. Например, если в корпусе 5 000 000 вхождений слов, а слово СОБАКА среди них встретилось 50 раз, то вероятность слова СОБАКА — одна стотысячная. Запишем это подобием формулы, где p означает вероятность:

$$p(\text{собака}) = \frac{\text{число вхождений слова СОБАКА}}{\text{число всех вхождений слов}} = \frac{50}{5\,000\,000} = \frac{1}{100\,000}$$

На основе корпуса можно заранее посчитать и запомнить вероятности всех слов. Тогда вероятность любого предложения (даже не встречавшегося в корпусе) можно посчитать, просто перемножив вероятности входящих в него слов, как это принято в теории вероятностей:

$$p(\text{птица сидит на крыше}) = p(\text{птица}) \cdot p(\text{сидит}) \cdot p(\text{на}) \cdot p(\text{крыше})$$

Этот способ тоже не очень хорош. Если слова переставить, вероятность не изменится. Но интуиция нам подсказывает, что вероятность предложения НА ПТИЦА КРЫШЕ СИДИТ должна быть ниже, чем ПТИЦА СИДИТ НА КРЫШЕ.

Преыдушие слова подсказывают нам, какими могут быть следующие слова, и это можно использовать при вычислении вероятностей. При этом, как показал 100 лет назад математик Андрей Марков, можно учитывать не весь предшествующий текст, а только несколько предыдущих слов. Если взять слишком большой предшествующий контекст, качество вычислений улучшится не слишком сильно, а сложность обработки возрастет.

Поэтому в популярных лингвистических моделях стали брать контекст из двух предыдущих слов. Так появилась триграммная модель языка. От слова «триграмма» — три идущих подряд слова.

Триграммная модель языка

Триграммная модель языка оценивает вероятности предложений. Для ее создания машине нужен большой корпус. Первым делом она соберет список входящих в него слов. Точнее, словоформ. Ведь СОБАКА и СОБАКОЙ — это разные последовательности символов, поэтому машина запомнит их как две разные единицы, не вникая, что за зверь за ними стоит, какая это часть речи и в какой грамматической форме. Важно только, что обе словоформы встретились в обучающем корпусе.

После этого для каждой тройки словоформ компьютер посчитает условную вероятность. Какова вероятность того, что после слов МНЕ ПОДАРИЛИ встретится слово СОБАКУ? Какова вероятность того, что после слов МНЕ ПОДАРИЛИ встретится слово НОСКИ? И все остальные варианты сочетаний. Условная вероятность записывается так:

р (собаку | мне подарили)
р (носки | мне подарили)
р (носков | мне подарили)
р (носками | мне подарили)

Справа от вертикальной черты указывается контекст МНЕ ПОДАРИЛИ. В предложении он предшествует слову СОБАКА, которое записывается слева от черты. Полная запись означает вероятность появления слова СОБАКА при условии, что у нас уже есть слова МНЕ ПОДАРИЛИ.

Если рассматривать все возможные тройки, то большинство сочетаний будут казаться бессмысленными. Какова вероятность того, что после слов СТАКАНЕ О встретится слово ГОРЕ?

р (горе | стакане о)

Однако это фрагмент из стихотворения Введенского:

и тишина была в стакане
о горе птичка говорит одна
не вижу солнечного я пятна
а мир без солнечных высоких пятен
и скуп и пуст и непонятен

На всякий случай надо посчитать вероятности для всех теоретически возможных троек слов. Отдельно считаются вероятности того, что слово встретится в самом начале предложения и перед ним ничего не будет, а также вероятность концов предложений во всех контекстах.

Все эти вероятности вычисляются на исходном корпусе и запоминаются. Они называются параметрами модели. После этого вероятность любой последовательности словоформ считается перемножением параметров:

$p(\text{птица сидит на крыше}) = p(\text{птица} | **) \cdot p(\text{сидит} | * \text{птица}) \cdot p(\text{на} | \text{птица сидит}) \cdot p(\text{крыше} | \text{сидит на}) \cdot p(\text{КОНЕЦ} | \text{на крыше})$

Последний параметр соответствует вероятности конца предложения после слов НА КРЫШЕ. Звездочки в двух первых параметрах означают начало предложения.

Осталось только понять, как на основе корпуса посчитать параметры модели — вероятности того, что слово встретится в том или в ином контексте. Машина может сделать это очень легко. Для $p(\text{собаку} | \text{мне подарили})$ она посчитает, сколько раз в корпусе встретился контекст МНЕ ПОДАРИЛИ. Например, 100 раз. А затем — сколько раз после этого контекста встретилось нужное слово СОБАКУ. Всего один раз. Значит, параметр — $1/100 = 0,01$. Конечно, чаще встретятся НОСКИ — раз 70. Параметр для них будет $70/100 = 0,7$.

$$p(\text{собаку} | \text{мне подарили}) = \frac{\text{число сочетаний МНЕ ПОДАРИЛИ СОБАКУ}}{\text{число сочетаний МНЕ ПОДАРИЛИ}} = \frac{1}{100} = 0,01$$

$$p(\text{носки} | \text{мне подарили}) = \frac{\text{число сочетаний МНЕ ПОДАРИЛИ НОСКИ}}{\text{число сочетаний МНЕ ПОДАРИЛИ}} = \frac{70}{100} = 0,7$$

Модель можно немного скорректировать («сгладить») и добавить математическую обработку случаев, которые не встретились в обучающем корпусе.

Машинное обучение и нейросети

В триграммной модели языка машина использует всего один признак: частоту появления слова после двух других слов.

Можно задавать машине другие признаки и решать другие задачи, используя математические методы. Например, можно взять корпус электронных писем, часть из них пометить как спам, а остальные — как не спам. Компьютер вычислит, с какой частотой и в каких контекстах встречаются слова типа скидка и распродажа, а затем научится находить спам среди новых сообщений. Похожим образом работает определение языка текста: компьютер сможет определять, что один абзац написан на английском, а другой — на испанском.

Классические алгоритмы машинного обучения хорошо справляются с подобными задачами классификации. Важно, чтобы у текстов были простые признаки. Например, характерные слова или знаки, типичные последовательности слов или повторяющаяся частотность. Нужно дать машине эти признаки, и она начнет анализировать данные.

Можно не размечать тексты заранее, но дать машине набор признаков. На их основании машина научится выделять похожие между собой элементы и объединять их в группы. Это называется кластеризация, и она используется, например, в контекстной рекламе или для рекомендации статей.

Но человек не всегда в состоянии найти подходящие признаки. Он хорошо замечает их только в самых очевидных случаях, но в более общих задачах, связанных с большим количеством признаков, начинает ошибаться. Как, например, отличить по стилю одного автора от другого? Какие признаки подобрать для машинного перевода? Наконец, как устроить сам алгоритм обработки больших массивов информации?

Для работы со сложными данными все чаще стали использовать нейронные сети. Именно они дают лучшие результаты в тех случаях, когда признаков так много, что непонятно, какие из них влияют на результат.

Нейронные сети возникли как попытка смоделировать на компьютере работу человеческого мозга. Мозг состоит из нейронов, каждый из которых принимает сигналы от других нейронов. Если уровень сигнала достаточно высок, нейрон передает его дальше.

Первый компьютер, моделирующий нейронную сеть, был создан в 1958 году американским ученым Фрэнком Розенблаттом. В его основе лежала простая модель, для которой Розенблатт придумал название **перцептрон**.

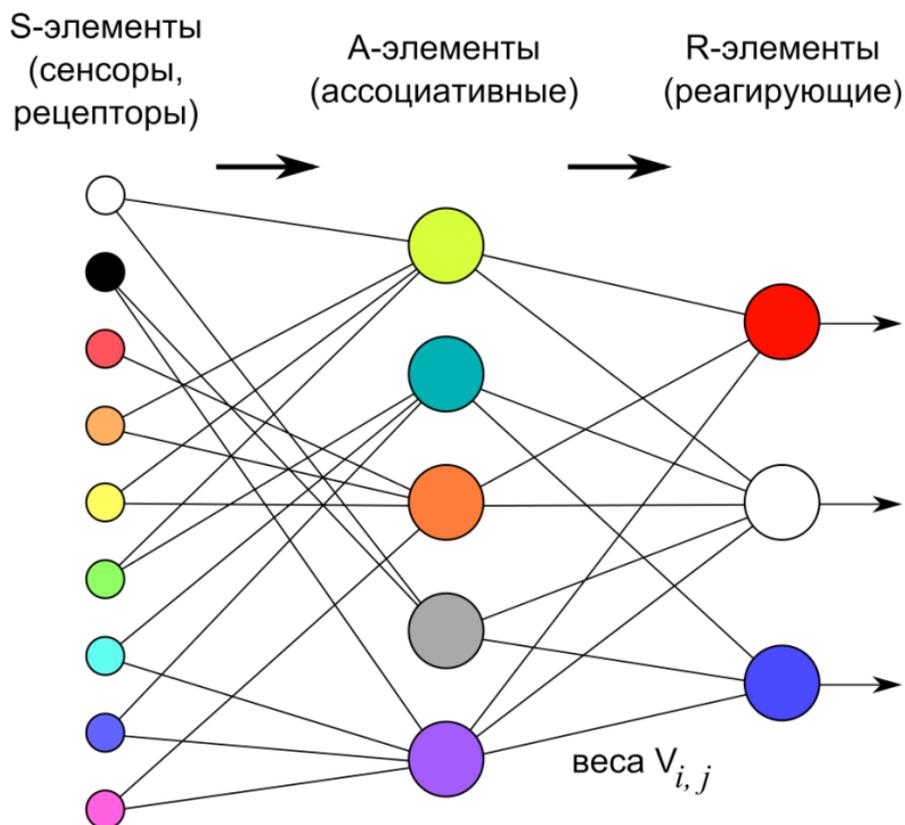


Рис. 3. Схема перцептрона Розенблатта
(Крайнов, А. (2009) *Общее представление перцептрона Розенблатта*.
URL: <https://commons.wikimedia.org/wiki/File:Perceptron-ru.svg>.
Изображение предоставлено по открытой лицензии)

Подробнее о том, как устроен перцептрон, будет рассказано в одном из следующих номеров журнала, пока приведем лишь очень краткое описание. Нейроны в перцептроне расположены слоями. Каждый нейрон передает сигнал только нейронам следующего слоя. У каждой передающей связи есть свой вес, и сигнал уменьшается или увеличивается пропорционально этому весу. Если вес 2, сигнал становится в два раза сильнее. А если он 0,1, то в десять раз слабее. Даже в жизни мы придаем большое значение сообщениям от некоторых людей, а слова других почти полностью пропускаем мимо ушей. Такая разная оценка сообщений приходит с жизненным опытом. Нейрону тоже надо подобрать правильный вес для каждой связи, и тогда на выходе получится нужный ответ.

Алгоритмы искусственных нейронных сетей совершенствовались, но принципиальный рывок произошел в 1974 году, когда А. И. Галушкин в МИФИ и Пол Вербос в Гарварде одновременно и независимо друг от друга предложили **метод обратного распространения ошибки**. Этот метод позволял машине настраивать веса, чтобы результат был все более точным. Вначале веса выбираются случайно. Затем по сети передается сигнал и оценивается ошибка на выходе. После этого, двигаясь по сети в обратном направлении, от последнего слоя к первому, нужно «подкручивать» веса так, чтобы ошибка становилась чуть меньше.

Затем пропускается новый обучающий сигнал; сравниваем результат с ответом и опять поправляем веса, двигаясь от последнего слоя к первому. Для этого алгоритма был разработан хороший математический аппарат. Так обучается сеть.

Количество слоев и типы связей — это архитектура нейросети. Архитектуру можно поменять. Например, разрешить нейрону передавать информацию самому себе, тогда сеть станет **рекуррентной**. Можно пропустить сеть через «бутылочное горлышко» — добавить слой с малым числом нейронов, чтобы собрать только самые важные признаки. Такой слой называется **сверточным**. Можно связать все нейроны со всеми, без всякого расслоения. Получится **цепь Маркова, машина Больцмана** или **сеть Хопфилда** — в зависимости от того, как нейрон обрабатывает входящие значения.

Разработчики и сами не всегда понимают, почему в каких-то задачах одна архитектура эффективнее другой. Они просто экспериментируют, а машина начинает гладко переводить, определять авторов текста или поддерживать беседу.

«Человек, не мешай!»

Итак, эволюция машинной обработки языка идет по пути «человек, не мешай!». Сначала компьютер избавляется от человеческих правил и переходит на машинное обучение, которое опирается на признаки, предложенные человеком. Затем он начинает находить нужные признаки самостоятельно, используя нейросети. Человек придумывает только архитектуру этих сетей. На следующем этапе, вероятно, оптимальную архитектуру тоже будет выбирать машина, но этому ее должен научить человек.

Далее возникает вопрос: сможет ли машина выйти на следующий уровень и без человека придумать архитектуру той сети, которая будет выбирать архитектуру другим сетям? А пойти еще дальше? Научится ли она обучаться всему вообще без человека? Ведь и у нас в голове как-то сам собой подбирается путь для установления нейронных связей. Станет ли такая система эффективнее, чем человеческий мозг?

Сейчас это часть большой философской проблемы, и современные философы разделяются тут на два примерно равных лагеря. Одни считают, что именно так и произойдет, потому что все должно моделироваться. Другие считают, что в человеческом восприятии всегда будет оставаться что-то невербализуемое, так называемые *qualia*, которые недоступны машине в принципе. Согласно представлениям этих философов, машина сможет смоделировать запах ландышей с химической стороны дела, но не сможет с главной — со стороны нашего ощущения от того, как это бывает, когда ты ощутил запах ландышей.

Но нейронные сети уже водят автомобиль лучше, чем человек. Они умеют генерировать на экране телевизора изображение диктора, который произносит заготовленный текст так искусно, что его не отличить от диктора-человека. Более того, если взять видеозапись вашего выступления, нейросеть сможет объединить ее с самым хулиганским в мире текстом и породит новую запись, на которой вы будете произносить этот текст своим голосом и в своей манере.

Все эти технологии появились буквально в самое последнее время. Эта научная область развивается так быстро, что учебники за ней не успевают. Но в Сети появляются курсы по созданию нейронных сетей. Лучшее, на наш взгляд, место для обучения нейросетям дистанционно — это Университет искусственного интеллекта (<https://neural-university.ru/>).

И бесспорно то, что будущее технологий в ближайшие десятилетия — за нейронными сетями. Открытие этих технологий изменит мир так же сильно, как изменило его появление компьютеров.

Сведения об авторе:

Ольга Владимировна Митренина, ORCID: [0000-0002-1750-5633](https://orcid.org/0000-0002-1750-5633), Web of Science ResearcherID: K-2876-2013, Scopus Author ID: 56558829300, e-mail: o.mitrenina@spbu.ru

Для цитирования: Митренина, О. В. (2019) Нейронные сети и компьютерная обработка языка. *Journal of Applied Linguistics and Lexicography*, 1 (2): 399–408.

Получена 28 августа 2019; **принята** 12 сентября 2019.

Права: © Автор (2019). Опубликовано Российским государственным педагогическим университетом им. А. И. Герцена. Открытый доступ на условиях лицензии CC BY-NC 4.0.

Author:

Olga V. Mitrenina, ORCID: [0000-0002-1750-5633](https://orcid.org/0000-0002-1750-5633), Web of Science ResearcherID: K-2876-2013, Scopus Author ID: 56558829300, e-mail: o.mitrenina@spbu.ru

For citation: Mitrenina, O. V. (2019) Artificial neural networks and natural language processing. *Journal of Applied Linguistics and Lexicography*, 1 (2): 399–408.

Received 28 August 2019; **accepted** 12 September 2019.

Copyright: © The Author (2019). Published by Herzen State Pedagogical University of Russia. Open access under CC BY-NC License 4.0.

УДК 81-112+83'373.6

ИСТОРИЯ ПИСЬМЕННОГО СЛОВА: О НЕКОТОРЫХ ПРИНЦИПАХ ЭТИМОЛОГИИ XVI В.

М. Л. Сергеев✉¹

¹ Российский государственный педагогический университет им. А. И. Герцена, 191186, Россия,
Санкт-Петербург, наб. реки Мойки, д. 48

HISTORY OF THE WRITTEN WORD: ON SOME PRINCIPLES OF 16TH CENTURY ETYMOLOGY

M. L. Sergeev✉¹

¹ Herzen State Pedagogical University of Russia, 48 Moika River Emb., Saint Petersburg 191186, Russia

Аннотация. В статье рассматриваются новые явления в практике этимологизирования, получившие распространение в XVI в. и оказавшие влияние на позднейшую историю лингвистики. Новшества были обусловлены значительным расширением языкового горизонта европейцев, появлением нефилологий и применением опыта классической филологии к изучению современных языков и их предьстории. Последнее обстоятельство нашло отражение, в частности, в поиске древнейших свидетельств по истории национальных языков в сочинениях римских авторов (Цезаря, Плиния Старшего, Тацита и др.). Встречающиеся в античных текстах германские или галльские глоссы, термины или имена собственные предположительно германского происхождения требовали этимологического объяснения. Этимон обнаруживался в лексике современных германских языков или известных филологам исторических формах (последовательное описание которых еще не велось), а формальные различия с засвидетельствованным у авторов словом объяснялись искажением последнего в рукописной традиции (при записи оригинального текста и, особенно, при создании его позднейших копий). Таким образом, в этимологии применялась филологическая аргументация, основанная на гуманистическом опыте издания и редактирования текстов, который, в частности, предполагал исправление испорченных мест в тексте с опорой на более надежные рукописные чтения (если они были доступны) или на конъектуры издателя. Использование этих принципов в этимологии XVI в. показано в статье на примере объяснения галльских имен на -rix и термина siloduni в работах швейцарских и немецких гуманистов (Э. Чуди, К. Гесснера и автора анонимного этимологического словаря германских имен собственных).

Abstract. The article covers some of the new tendencies in 16th century etymology, comparing them with the practice of ancient and medieval authors. These novelties were called forth by the expanding linguistic horizon of European scholarship, the emergence of “neo-philologies”, and the application of classical philological methods to the study of vernacular languages and their history. Remarkable evidence of the influence of humanist philology on linguistics is provided in 16th century reconstructions of the earliest testimonies of vernacular vocabulary in the heritage of Roman authors (i. e., Caesar, Pliny the Elder and Tacitus). Their texts contained a number of German and Gaulish “glosses”, as well as terms and names of presumably Germanic origin, which required etymological explanation. The “etyma” were sought in the vocabulary of modern Germanic languages and among historical forms available to scholars at the time, while the formal discrepancies between the reconstructed and the attested forms were often explained by the deteriorated state of written testimonies or circumstantial interferences that might have accompanied the composition, recording, and copying process of the literary text under consideration. Thus, etymologies relied on philological arguments based on the principles of humanistic editing, which included a collation of several manuscripts or emendation “ope ingenii” in the cases of text corruption. The author discusses this type of argumentation based on several examples of 16th century etymologies, namely the explanation of Gaulish personal names ending with -rix and the term “siloduni” by Swiss and German humanists (Aegidius Tschudi, Conrad Gessner and the author of the anonymous dictionary of German proper names).

Keywords: history of linguistics, history of philology, etymology, 16th century, humanism, German, Latin.

Ключевые слова: история языкознания, история филологии, этимология, XVI век, гуманисты, немецкий язык, латинский язык.

1. Этимология как инструмент познания и аргументации, а также как поэтический прием, была хорошо известна в античности и в Средние века; к этимологиям прибегали философы, историки, грамматик и поэты. Исследователи отмечают, что этимологические построения того времени, при всем их многообразии, были объединены преимущественным вниманием к соотношению имени и денотата и относительным пренебрежением к форме слова. Хотя изменения последней иногда классифицировали, вслед за Варроном (116–27 до н. э.), их не подчиняли определенным правилам, фактически они сводились к произвольным переменам и перестановкам букв/звуков. Еще одна важная особенность этимологий классического и средневекового периода в западноевропейской традиции состоит в том, что они были преимущественно ограничены латинским языковым материалом, хотя отдельные авторы обращались также к лексике древнегреческого и национальных языков (см.: Buridant 1998; Copeland, Sluiter 2009, 339–366; Sallmann 2004).

2. В некоторых отношениях ситуация заметно изменилась в эпоху Ренессанса. Первое изменение коснулось материала этимологий, второе было связано с развитием лингвистической и филологической теории. К середине XVI в. появилось позитивное осмысление многообразия языков, прежде ассоциировавшегося главным образом с вавилонским столпотворением и наказанием для человечества (Céard 1980); кроме того, под сомнение было поставлено использование по отношению к народным языкам термина «варварский язык», имевшего, несомненно, пейоративную коннотацию (ср. Rochette 1997). Причиной этих перемен стало, с одной стороны, значительное расширение языкового горизонта европейцев благодаря географическим открытиям и деятельности миссионеров; с другой стороны, деятельный интерес к древним и современным языкам — лингвистическое любопытство (ср. Considine 2017, 11–30) — стали проявлять ученые, занятые самыми разными областями науки, от богословской экзегезы до естественной истории и медицины (см. Sergeev 2018, 19–39).

3. Европейские «народные» языки вызывали двоякий интерес: богатство их словаря и стилистические возможности давали основания претендовать на статус литературного языка (ср. Burke 2004, 61–88); кроме того, представления об истоках языка и его родстве с другими языками оказывались важным инструментом для реконструкции пред истории народа, говорящего на нем. Национальная историография, подъем которой также пришелся на XVI век, опиралась на античные источники, образцы классической историографии и опыт их изучения гуманистами (см. Epenkel, Ottenheim 2017, 23–135). Работа историков предполагала обращение к языковым сопоставлениям и этимологиям — для объяснения иноязычных глосс и непонятных мест в античных свидетельствах, а также для подтверждения исторических реконструкций, касавшихся переселения народов, генеалогий и культурной преемственности. Так, немецкие и нидерландские авторы противопоставляли латинскому происхождению итальянского языка (и, соответственно, римским предкам итальянцев) обнаруженные сходства германского языка с греческим (ср. Dogonin 2016), который гуманисты считали более древним (Tavoni 1986). У французов, южные земли которых были когда-то колонизованы греками, были не меньшие претензии на греческое происхождение родного языка; более того, некоторые авторы полагали, что древние галлы (которых они считали предками французов) сами оказали влияние на греческий язык и науку (Dubois 1972, 47–50; Tavoni 1998, 50–55). В результате отдельные этимологии и в целом доля грецизмов

в языке стали предметом оживленных споров германоязычных и франкоязычных филологов (ср., например, оценку А. Юния: Junius 1556, 207–210).

4. Что касается развития лингвистической теории, то в XVI–XVII вв. постепенно были сформулированы три важнейших основания сопоставления языков, в дальнейшем использованные в сравнительно-историческом языкознании: разграничение заимствованной и унаследованной лексики, определение морфологического состава сближаемых слов в разных языках и установление регулярных фонетических соответствий (Droixhe 1984; Muller 1984; Muller 1986). Так, например, Ш. де Бовелль в «Книге о различии народных языков» показывает соответствие лат. *b* ~ франц. *v*, ссылаясь на пары *habere* ~ *avoir* ('иметь'), *ebur* ~ *uvoire* ('слоновая кость'), *ebrius* ~ *uvre* ('пьяный') (Bovelles 1533, 26). Юст Липсий предваряет список германо-персидских соответствий замечанием о том, что среди них есть явные заимствования из латыни, вроде названий вина, оливкового масла, смоквы и др. (Lipsius 1614, 58–59). Однако на практику этимологизирования в рассматриваемую эпоху формальные правила оказывали скорее спорадическое влияние, применялись *ad hoc*, чтобы быть нарушенными уже в соседних примерах.

5. Вместе с тем представляется, что в этимологических построениях гуманистов заметен новый способ аргументации, появлению которого лингвистика была обязана развитию филологической критики текста. Практика издания античных авторов к началу XVI в. подразумевала исправление непонятных мест текста, как с помощью конъектур, так и учитывая разночтения в доступных рукописях, особенно тех, которые считались древними (ср. Nellen, Bloemendal 2014). Работа издателя и редактора книги состояла, по существу, в поисках «этимона» текста, дошедшего в поздних списках — «несовершенных», «искаженных», «испорченных».¹ Как видно, у филологической деятельности было много общего с этимологической реконструкцией, стремившейся устранить последствия «порчи» слов, произошедшей с течением времени.² Более того, иногда филологам приходилось прибегать к этимологиям для установления правильной формы слов в изучаемых текстах. Особенно это касалось случаев, когда латинский или греческий текст включал иноязычный (или предположительно иноязычный) материал, например, галльские имена собственные или германские глоссы.

6. Для интерпретации или исправления такого рода слов (как правило, подвергшихся определенной фонетической и морфологической адаптации к латинскому тексту) немецкие гуманисты охотно прибегали к поиску этимонов в родном языке. При этом, следуя принципам реконструкции письменного источника, они часто объясняли формальные различия между засвидетельствованным в латинской рукописи словом и предполагаемым немецким соответствием исключительно превратностями истории текста. Объективные изменения в самом германском языке, которые должны были произойти за более чем тысячу лет, прошедших с момента написания рассматриваемых книг, не принимались во внимание: в значительной степени это было обусловлено недостаточной изученностью средневековых памятников (ср. Kibbee 1992). Далее будет рассмотрено несколько этимологий из сочинений XVI в., для подкрепления которых авторы использовали филологическую аргументацию; все этимологизируемые имена были взяты из «Записок о Галльской войне» Цезаря.

¹ «Codices imperfecti, depravati, corrupti»: (см. Rizzo 1973, 221–226).

² Так языковые изменения трактовались уже в работах античных филологов (см. Müller 2003).

(1) В популярном словаре древнегерманских имен собственных «*Aliquot nomina propria Germanorum ad priscam etymologiam restituta*» (1537), авторство которого приписывали М. Лютеру, имя кельтского вождя Верцингеторикса³ производится от «саксонского» словосочетания «*Hertoge Hinric*» («герцог Генрих»). Превращение гипотетического **hertoge Hinric* в засвидетельствованное в текстах *Vercingetorix* объясняется следующим образом: «по вине писцов были перепутаны, переставлены и испорчены буквы; ведь он [Цезарь] хотел записать саксонское *Hertoge Hinric*, то есть герцог Генрих, а позже писцы переменили Н на V, переставили *toge* после *Hin* и превратили в *geto*».⁴ Манипуляции, произведенные с элементами слова (замена букв, метатеза), были хорошо знакомы этимологам от античности до XVIII в.; менее привычным кажется их объяснение: изменение формы слова не подается как нечто само собой разумеющееся и не приписывается неверному произношению («*vitium labiorum*»: ср. *Bovelles* 1533, 28, 81, 91), но связывается с ошибками писцов, то есть локализовано в скриптории.

(2) Швейцарский историк Эгидий Чуди (1505–1572) в сочинении о древностях Альпийской Реции приводит ряд аргументов в пользу того, что древний галльский язык должен считаться германским диалектом⁵, в том числе обращается к этимологиям галльских имен собственных — непременно инструменту в решении этого вопроса. В качестве первого примера Чуди рассматривает компонент *-rix* (в таких именах, как *Orgetorix*, *Dumponix*, *Ambiorix*, упоминавшееся выше *Vercingetorix* и т. д.), который он отождествляет с немецким *-rich* (в именах *Friderich*, *Henrich*, *Dietrich* и др.). Превращение «*ch*» в «*x*» объясняется им графически и, как и в предыдущем эпизоде, связывается с историей конкретного литературного памятника, но здесь виновником искажения формы слова называется сам автор (выступивший в роли переписчика). По версии Чуди, «Цезарь обнаружил имена гельветов, записанные греческими буквами⁶, некоторые из которых заканчивались на греческую букву χ , которая передает [немецкое] *ch*, так как очень многие германские имена имеют такое окончание, каковы Фридрих, Генрих, Дитрих, Адельрих, Ульрих и т. д.», но поскольку латинский язык не допускает *-ch* на конце слова, «Цезарь сделал из греческой буквы χ латинскую *x*...»⁷ (вероятно, учитывая их внешнее сходство). Объяснение показалось автору настолько удачным, что он не задался очевидным вопросом: почему римлянин, долгое время воевавший в Галлии и прекрасно знавший имена галлов, должен был довериться прочтению (даже если имена, перечисленные Чуди, действительно были бы записаны в гельветских табличках таким образом), а не своим знаниям и слуху? Чуди, таким образом, представляет Цезаря в образе современного ему гуманиста, который знал о галлах только из письменных свидетельств, с пиететом относился к древнегреческим текстам и хорошо усвоил правила латинской грамматики, определяющие, какие «окончания» слов допустимы в именительном падеже, а какие — нет. Следует заметить, что этимология Чуди была охотно принята другими немецкими авторами: ее воспроизвел Г. Глареан в комментарии к Цезарю (*Caesar* 1544, 48), а затем К. Гесснер в «МитриDATE» (*Gessner* 1555, 22b).

³ Современная кельтология объясняет имя *Vercingetorix* как трехчастный композит (*Ver-cingeto-rix*), значащий «верховный правитель воинов» (см. *Delamarre* 2003, 116, 260–261, 314).

⁴ «*Sed scriptorum vitio confusis, transpositis & corruptis literis, Voluit enim Saxonicum illud, Hertoge Hinric scribere, id est, Dux Henricus, Et postea scriptores mutarunt H in V. & transposuerunt toge, post Hin & geto fecerunt*» (*Aliquot nomina* 1537, B1b).

⁵ Наиболее распространенная версия о природе галльского языка у немецких авторов XVI–XVII вв. (см. *Sergeev* 2011; *Van Hal* 2013–2014).

⁶ Ср. «Записки о Галльской войне» (1, 29) и комментарий Э. Чуди (*Tschudi* 1538, 105).

⁷ «*Caesar cum reperisset Helvetiorum virorum nomina Graecis literis conscripta, quaedam desinebant in χ Graecam literam, quae facit ch, eo quod plurima Germanica nomina hanc habeant finalem terminationem, qualia sunt Friderich, Henrich, Dietrich, Adelrich, Ulrich etc. quam finalem syllabam latinus sermo ferre non potest, & propterea fecit Caesar ex Graeco χ literam latinam *x*...» (*Tschudi* 1538, 119–120).*

(3) В справочнике о языках мира «Митридат» (1555) в главе «De lingua Germanica» К. Гесснер (1516–1565) рассматривает ряд терминов, приведенных Цезарем и Плинием и имевших, по его предположению, германское происхождение (Gessner 1555, 32b–34a). В параграфе, озаглавленном в типографской маргиналии «Siloduni vel Solidurii», приводится сообщение Николая Дамаскина (I в. до н. э.)⁸ о том, что у вождя одного из кельтских племен был отряд отборных воинов, которых галаты <sic!> на своем языке (sermone patrio) называли силодунами (Silodunos). Гесснер полагает, что Николай эти сведения заимствовал у Цезаря⁹, который в «Записках о Галльской войне» (3.22) рассказывает о «солдуриях» (soldurii) — преданных воинах вождя аквитанского племени. Чтобы согласовать два свидетельства, Гесснер отождествляет упоминаемые в них термины, точнее, выводит один из другого: «Цезарь <...> называет их не солидунами <sic!>, а солидуриями, некоторые читают в три слога — солдурьи».¹⁰ Таким образом, предполагается следующая последовательность форм (обратная порядку их появления в тексте), объясняющая «силодунов» у Николая Дамаскина: siloduni < *soliduni < solidurii, soldurii. Соотношение между формами “solidurii” и “soldurii” Гесснер описывает филологической формулировкой «некоторые читают [solidurii] в три слога [как soldurii]» (quidam tribus syllabis legunt). Однако первый вариант — «solidurii» — мне не удалось обнаружить ни в изданиях XVI в., ни в критическом аппарате. Возможно, Гесснер предположил существование утраченного чтения «solidurii», из которого затем появилась столь же гипотетическая форма «soliduni», в результате метатезы якобы превратившаяся в «siloduni» у Дамаскина. Далее термин, засвидетельствованный в тексте Цезаря, получает этимологическое объяснение: Гесснер производит его из нем. Sold(um) ‘плата (за военную службу)’, Söldner ‘(солдаты) служащие за плату’ или из нем. soll(en) ‘долженствовать’ и duren (durten) ‘отваживаться, осмеливаться’. Латинская этимология — «(a) solide durando», то есть «от твердой выносливости», — отвергается, так как «известно, что это слово галльское или германское».¹¹ Мы вновь видим связь этимологии с филологической аргументацией: в данном случае этимология подтверждает первичность формы слова, засвидетельствованной в тексте Цезаря (по отношению к той, что у Дамаскина), а предполагаемое разночтение облегчает реконструируемый «переход» soldurii в siloduni.

Рассмотренные примеры не уникальны и при всей курьезности предложенных реконструкций показывают существенные изменения в этимологической аргументации, появляющиеся в XVI в. Для объяснения предполагаемых соответствий между словами одного или нескольких языков недостаточным оказывается указание на общую семантику слов или на их отдаленное созвучие: авторы ищут дополнительный лингвистический материал и обоснования постулируемых формальных переходов.¹² В эпоху возрождения классической словесности, издания греческих и латинских авторов источником такого обоснования становится *история текста*, реальная или воображаемая.

⁸ Сохранившееся в цитате в «Пире мудрецов» Афиня (6, 249a-b).

⁹ «Unde Nicolaus sua forte transtulit» (Gessner 1555, 33b).

¹⁰ «Hos quidem C. Caesar <...> non solidunos, sed solidurios vocat, quidam tribus syllabis soldurios legunt» (Gessner 1555, 33b).

¹¹ «Dictionem Gallicam seu Germanicam esse» (Gessner 1555, 33b). Примечательно, что нем. Sold происходит как раз из лат. solidus ‘твердый, прочный’, использовавшегося как обозначение монеты (см. Deutsches Wörterbuch, Bd. 16, 1433–1434).

¹² Этому аспекту истории этимологической аргументации в XVI–XVII вв. почти не уделяется внимания в исследованиях (ср. Hassler 2009).

Sources

- Aliquot nomina propria Germanorum ad priscam etymologiam restituta.* (1537) Vitembergae: s. n., [16] fol. (In Latin)
- Bovelles, C. (1533) *Liber de differentia vulgarium linguarum, & Gallici sermonis varietate* <...>. Parisiis: R. Stephanus, 107, [1] p. (In Latin)
- Caesar, C. (1544) *C. Iulii Caesaris commentariorum libri VIII. Quibus adiecimus suis in locis D. Henrici Glareani doctissimas annotationes.* Basileae: N. Brylinger, [32], 741, [41] p. (In Latin)
- Dubois, J. (1531) *In linguam Gallicam Isagoge, una cum eiusdem Grammatica Latinogallica, ex Hebraeis, Graecis, et Latinis authoribus.* Parisiis: R. Stephanus, [16], 159, [1] p. (In Latin)
- Gessner, C. (1555) *Mithridates. De differentiis linguarum tum veterum tum quae hodie apud diversas nationes in toto orbe terrarum in usu sunt. Tigurini observationes.* Tiguri: Excudebat Froschoverus, [2], 78 fol. (In Latin)
- Junius, H. (1556) *Animadversorum libri sex, omnigenae lectionis thesaurus, in quibus infiniti pene autorum loci corriguntur et declarantur, nunc primum et nati, et in lucem aediti.* <...> Basileae: M. Isengrin, [56], 432 p. (In Latin)
- Lipsius, J. (1614) *Iusti Lipsii Epistolarum selectarum centuria tertia ad Belgas.* Antverpiae: ex officina Plantiniana, [4] fol., 118 p. (In Latin)
- Tschudi, A. (1538) *De prisca ac uera Alpina Rhaetia, cum caetero Alpinarum gentium tractu, nobilis ac erudita ex optimis quibusque ac probatissimis autoribus descriptio.* Basileae: Apud Mich. Isengrin, [8], 134 p. (In Latin)

Dictionaries

- Delamarre, X. (2003) *Dictionnaire de la langue gauloise: Une approche linguistique du vieux-celtique continental.* 2 éd. Paris: Errance, 440 p. (In French)
- Deutsches Wörterbuch von Jacob Grimm und Wilhelm Grimm.* [Online]. Available at: <http://woerterbuchnetz.de/DWB/> (accessed 07.06.2019). (In German)

References

- Buridant, C. (1998) Les paramètres de l'étymologie médiévale. In: *L'étymologie de l'antiquité à la renaissance.* Villeneuve-d'Ascq: Presses universitaires du Septentrion, pp. 11–56. (In French)
- Burke, P. (2004) *Languages and communities in early modern Europe.* Cambridge: Cambridge University Press, XIV, 210 p. (In English)
- Céard, J. (1980) De Babel à la Pentecôte: La transformation du mythe de la confusion des langues au XVIe siècle. *Bibliothèque d'Humanisme et Renaissance*, 42 (3): 577–594. (In French)
- Considine, J. (2017) *Small dictionaries and curiosity: Lexicography and fieldwork in Post-Medieval Europe.* Oxford: Oxford University Press, 336 p.
- Copeland, R., Sluiter, I. (eds.). (2009) *Medieval grammar and rhetoric: Language arts and literary theory, AD 300–1475.* Oxford: Oxford University Press, X, 972 p. (In English)
- Doronin, A. V. (2016) Brat'ya Tuiskon i Gomer, druidy i pes abbata Tritemiya: kak nemetskie gumanisty rodnilis' s drevnimi grekami. In: O. F. Kudryavtsev (ed.). *Iskusstvo i kul'tura Evropy epokhi Vozrozhdeniya i rannego Novogo vremeni.* Moscow; Saint Petersburg: Tsentr gumanitarnykh initsiativ Publ., pp. 269–289. (In Russian)
- Droixhe, D. (1984) *Avant-propos. Histoire Épistémologie Langage*, 6 (2): 5–16. (In French)
- Dubois, C.-G. (1972) *Celtes et gaulois au XVIe siècle: Le développement littéraire d'un mythe nationaliste.* Paris: Librairie philosophique J. Vrin, 205 p. (In French)
- Enenkel, K. A. E., Ottenheim, K. (2017) *Oudheid als ambitie: De zoektocht naar een passend verleden 1400–1700.* Nijmegen: Vantilt, 349 p. (In Dutch)
- Haßler, G. (2009) Etymologie. In: G. Haßler, C. Neis. (Hrsg.). *Lexikon sprachtheoretischer Grundbegriffe des 17. und 18. Jahrhunderts.* Bd. 1. Berlin: De Gruyter, S. 625–658.
- Kibbee, D. A. (1992) Renaissance notions of medieval language and the development of historical linguistics. *The journal of medieval and renaissance studies*, 22 (1): 41–54. (In English)
- Muller, J.-C. (1984) Quelques repères pour l'histoire de la notion de vocabulaire de base dans le précomparatisme. *Histoire Épistémologie Langage*, 6 (2): 37–43. (In French)
- Muller, J.-C. (1986) Early stages of language comparison from Sasseti to Sir William Jones (1786). *Kratylos*, 31: 1–31. (In English)
- Müller, R. (2003) Konzeptionen des Sprachwandels in der Antike. *Hermes*, 131 (2): 196–221. (In German)

- Nellen, H. J. M., Bloemendal, J. (2014) Philology: Editions and editorial practices in the early modern period. In: Ph. Ford, J. Bloemendal, Ch. Fantazzi (eds.). *Brill's Encyclopaedia of the Neo-Latin World*. Leiden: Brill, pp. 185–206. (In English)
- Rizzo, S. (1973) *Il lessico filologico degli umanisti*. Roma: Edizioni di Storia e Letteratura, XXIV, 394 p. (In Italian)
- Rochette, B. (1997) Grecs, Romains et Barbares: À la recherche de l'identité ethnique et linguistique des Grecs et des Romains. *Revue belge de Philologie et d'Histoire*, 75 (1): 37–57. (In French)
- Sallmann, K. (2004) Etymology. In: H. Schneider, H. Cancik (eds.). *Brill's new Pauly*. Vol. 5. Leiden: Brill, columns 123–126. (In English)
- Sergeev, M. L. (2011) Kommentarij k spisku "gall'skikh" slov v "Mitridate" K. Gessnera [The list of "Gaulish" words in C. Gessner's "Mithridates": A commentary]. In: N. A. Bondarko, N. N. Kazanskij (eds.). *Acta linguistica Petropolitana*. Vol. 7. Pt. 1. Saint Petersburg: Nauka Publ., pp. 386–408. (In Russian)
- Sergeev, M. L. (2018) *Sopostavlenie yazykov v XVI veke (na primere "Mitridata" (1555) Konrada Gessnera)*. PhD dissertation (Philology). Saint Petersburg, Institute for Linguistic Studies of the Russian Academy of Sciences, 234 p. (In Russian)
- Tavoni, M. (1986) On the Renaissance idea that Latin derives from Greek. *Annali della Scuola normale superiore di Pisa. Serie III*, 16 (1): 205–238. (In English)
- Tavoni, M. (1998) Renaissance linguistics: Western Europe. In: G. Lepschy (ed.). *History of linguistics. Vol. 3: Renaissance and Early Modern linguistics*. London; New York: Longman, pp. 1–108. (In English)
- Van Hal, T. (2013–2014) From *Alauda* to *Zythus*. The emergence and uses of Old-Gaulish word lists in early modern publications. In: *Keltische Forschungen*. Bd. 6. Wien: Praesens Verlag, S. 219–277. (In English)

Сведения об авторе:

Михаил Львович Сергеев, ORCID: [0000-0002-1548-3901](https://orcid.org/0000-0002-1548-3901), e-mail: librorumcustos@gmail.com

Для цитирования: Сергеев, М. Л. (2019) История письменного слова: о некоторых принципах этимологии XVI в. *Journal of Applied Linguistics and Lexicography*, 1 (2): 409–415.

Получена 3 июля 2019; **принята** 21 июля 2019.

Права: © Автор (2019). Опубликовано Российским государственным педагогическим университетом им. А. И. Герцена. Открытый доступ на условиях лицензии CC BY-NC 4.0.

Author:

Mikhail L. Sergeev, ORCID: [0000-0002-1548-3901](https://orcid.org/0000-0002-1548-3901), e-mail: librorumcustos@gmail.com

For citation: Sergeev, M. L. (2019) History of the written word: On some principles of 16th century etymology. *Journal of Applied Linguistics and Lexicography*, 1 (2): 409–415.

Received 3 July 2019; **accepted** 21 July 2019.

Copyright: © The Author (2019). Published by Herzen State Pedagogical University of Russia. Open access under CC BY-NC License 4.0.