Data-Driven Approach to Identification of Latin Phrases in Russian Web-Crawled Corpora

Vladimír Benko[1,2] and Katarína Rausová[1]

[1] Slovak Academy of Sciences, Ľ. Štúr Institute of Linguistics, Bratislava, Slovakia
[2] Comenius University in Bratislava, UNESCO Chair in Plurilingual and Multicultural Communication, Bratislava, Slovakia

{vladimir.benko,katarina.rausova}@juls.savba.sk

Abstract

Latin phrases are an integral part of the language of educated speakers in many (European) languages. Besides lexical units of Latin origin that have been already adapted to the orthography of the respective host language and calques, phrases retaining the original form and orthography can also be found in many texts. Due to the rather low frequency of the phenomenon, however, any systematic attempt of its analysis was a real challenge before the advent of very large (multi-Gigaword) corpora.

Our paper presents a method of semi-automatic detection of Latin phrases in a Russian web corpus based on applying a Latin tagger and a series of filtrations performed by standard Linux utilities. The preliminary analysis of the resulting candidate list is shown in the concluding part of the paper.

Keywords: Latin Quotations, Code Switching, Corpus-Driven Approach

Reference for citation: Benko V., Rausová K. Data-Driven Approach to Identification of Latin Phrases in Russian Web-Crawled Corpora // Computer Linguistics and Computing Ontologies. Vol. 4 (Proceedings of the XXIII International Joint Scientific Conference «Internet and Modern Society», IMS-2020, St. Petersburg, June 17-20, 2020). - St. Petersburg: ITMO University, 2020. P. 11 – 20. DOI: 10.17586/0000-0000-2020-4-11-20

> *Более того, здесь есть своя "старуха ex machina" — Антонида Васильевна, внезапно возвращающаяся с порога смерти, меняющая расклад в семействе Загорянских и заражающая главного героя — учителя Алексея — игорной страстью [1].*

Introduction

Amount of lexical evidence for low-frequency lexical items, such as idioms and other types of fixed expressions, could hardly be considered sufficient not only in the pre-corpus times, but also during early decades of corpus linguistics. Linguistic analysis of and lexicographic treatment of such phenomena had to be based on a rather small number of examples found in collections of citations slips, or often hapax occurrences in first-generation corpora. Even with a 100 Megaword corpus at hand, a corpus-based methodology only could be applied, i.e. attesting occurrences of the "suspected" phrases based on their lists found in legacy lexicographic works.

With the advent of the "big data" paradigm to corpus linguistics in the form of multi-Gigaword corpora, as well as with the availability of robust tolls for their linguistic annotation, the situation gradually began to change. Russian also belongs to languages with corpora of this class available, such as enTenTen [2], GICR [3, 4], Araneum Russicum [5], Taiga [6], or Omnia Russica [7]. Having such resources at hand, linguists are not only capable of finding many more occurrences of the phrases their existence was known in advance but can also try to apply a potentially more efficient – corpus-driven – approach to identifying and analyzing them in the corpus. Our paper presents an experiment in the framework of which we tried to apply such a data-driven approach to semi-automatic identification of Latin text fragments in a multi-Gigaword Russian corpus. Due to its availability in the source format, we decided to use the data of Araneum Russicum Maximum [8], a web-crawled corpus developed in the framework of the Aranea Project [9].

1. Latin Phrases in Russian Texts

Within the Russian linguistic tradition, Latin quotations (or other types of Latin text chunks) appearing in Russian texts are usually referred to as крылатые выражения (“winged expressions”), a term originating back in antique times (it can be found in Homer’s Odyssey). A similar term is used in some other languages, such as German (“geflügelte Wörter”), or Polish (“skrzydlate słowa”). On the other hand, this term is rarely used in English linguistics (though it can be found in some sources [10]), and all foreign language text fragments are usually described as quotations. From the contrastive linguist’s perspective, this situation may be somewhat confusing, as both terms cannot always be treated as synonyms.

At this stage of our work, we were only interested in expressions written in Latin script appearing in Russian corpora (i.e. not only those falling to “quotation” category), though some other types of Latin-originated expressions can also be found in Russian texts, for example, phrases transliterated to Cyrillics (e.g. де-факто, де-юре).

2. Related Work

As far as we were able to find out, the phenomenon of Latinisms appearing in Russian texts is usually studied from the “Latin origin perspective” only, i.e. research papers [11] and lexicographic works [12] concentrate on Latin phrases translated to Russian.

Table 1. Latin phrases in the Russian National Corpus [16]

	Latin expression	**Fiction**	**Non-fiction**
Group I	etc	95	1165
	nota bene	14	11
	persona non grata	2	4
	post factum	5	42
	pro et contra	2	28
	Subtotal	**118**	**1250**
Group II	De gustibus non est disputandum	5	0
	Fiat lux!	1	2
	Memento mori	32	33
	Per aspera ad astra	3	7
	Urbi et orbi	2	33
	Subtotal	**43**	**75**
Group III	Alea jacta est	12	5
	In vino veritas	10	4
	O tempora, o mores!	2	4
	Panem et circenses	3	2
	Veni, vidi, vici	5	5
	Subtotal	**32**	**20**
	Total	**193**	**1345**

Among notable exceptions, we can find works of Grudeva [13] (also with Pavlova [14]) who studied appearance of a collection of 15 Latin expressions in the Russian National Corpus

(RNC) [15]. She divided them into three lexical groups (clichés, proverbial expressions, quotations), compared their frequencies in fiction and non-fiction texts [16] – see Table 1 – and also provided a breakdown by the time of origin of the respective texts (not shown here).

These results are not easy to fully reproduce today, as the size of the RNC has slightly increased meanwhile. But even today's frequencies are not significantly higher, with over 85% of all occurrences represented by the same single expression (*etc*). It is therefore obvious that RNC can hardly be regarded as a sufficient resource for studying of Latinisms in Russian texts.

3. Needles in a Haystack

Russian corpora naturally contain some amount of non-Cyrillic tokens, consisting mostly of letters of the Latin alphabet. They may represent not only words coming from various languages (predominantly English, German and French, even transliterated Russian), but also all sorts of proper names (e.g. Google, Yandex, iPhone), acronyms and abbreviations (HDMI, BMW, inc.), Roman numerals (XXI, viii), physical units (Kbps), variables in equations, URLs and e-mail addresses, etc. Within the context of our work, we are mostly interested only in all-Latin tokens [17] that are potential candidates for Latinisms. Table 2 shows the situation in three Araneum Russicum III web corpora of different sizes. The figures were extracted directly by querying the respective corpora via the web interface of the NoSketch Engine [18] corpus manager [19].

Table 2. Latin characters in the Russian Aranea corpora

	Minus (125M)		**Maius (1.25G)**		**Maximum (19.8G)**	
Tokens (with)...	hits	i.p.m.	hits	i.p.m.	hits	i.p.m.
...at least one Latin char	1,256K	10,049	12,751K	10,201	201,603K	10,193
...all-Latin	997K	7,972	10,127K	8,102	160,297K	8,104
...at least two chars long	936K	7,490	9,528K	7,623	150,651K	7,617
...lowercase-only	168K	1,343	1,770K	1,416	28,012K	1,416

It can be seen from the table that (1) the relative frequencies (i.p.m.) change only insignificantly with the change corpus size, and (2) even if we only considered the all-lowercase candidate strings from the smallest corpus, their manual analysis would hardly be feasible.

3.1. A "Brute Force" Approach

Fortunately, we do not have to look for potential Latinisms manually and can use a powerful tool – the Latin tagger. Surprisingly enough, Latin does not belong to low-resourced languages [20] – several treebanks are available that enabled the creation of language models for (to our knowledge) at least two taggers with a FLOSS license. TreeTagger [21] comes with two different language models that were trained on various corpora [22], while UDPipe [23] works with a model [24] trained on one of the Latin treebanks developed within the Universal Dependencies [25] Project. More suitable for our work is TreeTagger, due to one of its key features – it provides for explicit indication of the out-of-vocabulary (OOV) lexical items. All word forms not found during the morphological lexicon lookup are flagged as "<unknown>", which will be the tagging result of the bulk of the data in a Russian corpus.

To make the processing as simple as possible, we decided not to care about the efficiency by attempting to optimize the processing time, and used mostly the standard Linux utilities (such as grep, sort, uniq, etc.)

3.2. The Method

The overall idea of our quest for Latin phrases can be summarized as follows:

1. Take the tokenized Russian corpus, delete annotation (if any).
2. Run a Latin tagger on it.

3. Delete tokens tagged as <unknown> (leaving empty lines).
4. Delete numbers and punctuation.
5. Delete annotation (tags and lemmas).
6. Merge multiple empty lines.
7. Change newlines after consecutive non-empty lines to spaces (i.e. putting multi-word expressions at the same line).
8. Produce a frequency list.

Some of these steps applied to a Russian sentence [26] (producing the “de facto” expression) are shown in Table 3.

Table 3. Processing steps 2, 3 and 4 + 5

Word	**Tag**	**Lemma**	**Word**	**Tag**	**Lemma**	**Word**
Но	N:voc	<unknown>				
если	N:voc	<unknown>				
мы	NPR	<unknown>				
будем	N:voc	<unknown>				
говорить	NPR	<unknown>				
о	N:voc	<unknown>				
положении	NPR	<unknown>				
de	**PREP**	**de**	**de**	**PREP**	**de**	**de**
facto	**N:abl**	**factum**	**facto**	**N:abl**	**factum**	**facto**
,	PUN	,	,	PUN	,	
то	N:voc	<unknown>				
есть	N:voc	<unknown>				
,	PUN	,	,	PUN	,	
грубо	N:voc	<unknown>				
говоря	N:voc	<unknown>				
,	PUN	,	,	PUN	,	
два	N:voc	<unknown>				
варианта	NPR	<unknown>				
.	SENT	.	.	SENT	.	

It is obvious, that most sentences (or even entire documents) will *not contain any* all-Latin tokens. However, the deduplication of empty lines (by means of the Linux *uniq* utility) makes the problem of ignoring the non-Latin (mostly Cyrillic) tokens really simple.

3.3. The Processing

The speed of processing naturally depends not only on the amount of data but also on the computing power available – a multi-core machine with plenty of main memory is definitely a great advantage here so that the most compute-intensive operation – tagging – could be run in several parallel processes. The whole procedure is shown in Table 4.

Table 4. Processing times

Operation	**Tools used**	**Elapsed time**
Deleting the original annotation	*cut* utility	2:54:34
Splitting source vertical to 10 parts	custom splitter	1:01:20
Tagging (10 parallel processes)	*TreeTagger*	9:07:21
Merging the tagged data	*cat* utility	2:48:56
Removing <unknown> and punctuation tokens, merging empty lines, deleting lemma a tag	*egrep*, *uniq* and *cut* utilities	8:15:14
Producing the frequency distribution	*sort* and *uniq* utilities	0:00:02
Total time elapsed		**24:07:27**

Taking into account the huge amount of data to be processed (size of the source file was 644.8 GiB, yielding 258.2 GiB after removing the original annotation), the overall processing time is not that surprising. The most striking information is that, though it took slightly more than one day to compute the “rough” candidate list, the resulting distribution could finally be processed just in two seconds.

4. Preliminary Results

Out of the 19,778,053,615 tokens of Araneum Russicum III Maximum, the procedure described above produced a list of 240,607 different candidate expressions consisting of two or more words, with 82,817 (34.4%) of them having a non-hapax frequency. It is obvious that even such a list is too large for a manual analysis.

Table 5. Most frequent candidate expressions

Freq	Expression
21853	*Read more*
15881	***in vitro***
8558	***Homo sapiens***
7671	***in vivo***
7550	*Made in*
6758	*Canon EOS*
5439	*read more*
4825	*Read More*
4782	***homo sapiens***
4423	*Credit Suisse*
4264	*made in*
4187	*Chrome OS*
3819	***in situ***
3436	*Institute for*
3203	*PS Vita*
2951	*ID NO*
2890	***Deus Ex***
2885	*TRACE MODE*
2824	***de facto***
2756	***ad hoc***
2748	*de Paris*
2692	***Homo Sapiens***
2560	***status quo***
2439	*Super Mario*
2352	*Lotus Notes*
2250	***alma mater***
2096	*Face ID*
2082	***Candida albicans***
1959	*Liqui Moly*
1928	***Homo erectus***
1802	*Junior Suite*
1773	*are here*
1627	*Da Vinci*
1604	*EOS for*
1506	*LIQUI MOLY*
1454	*Marco Polo*
1444	***Opus Dei***
1439	*it is*
1434	*in der*
1409	*Video Editor*
1408	***terra incognita***
1406	***Et Cetera***
1344	*et de*
1318	*do it*
1282	***alter ego***
1276	*Creative Suite*
1267	*Athlon II*
1252	***per os***
1251	*Note II*
1244	*Focus ST*
1204	*Opera Mobile*
1204	*It is*
1173	*Jerusalem Post*
1141	***Natus Vincere***
1141	***Alma Mater***
1122	***Homo habilis***
1095	*Lotus Domino*
1093	*SATA III*
1092	*Do It*
1065	*Runa Capital*

But let us first have a look at the data before attempting any further (semi-)automated processing. Table 5 shows sixty most frequent items of the list.

Some items in the table look surprising, or even amusing. The “*Read more*” expression heading the list was (nonsensically) analyzed as consisting of two Latin nouns, and tagged as ***Read/N:abl/rea*** and ***more/N:abl/morus|mos***, i.e. “*culprit*, *sinner*” and “*black mulberry tree*” ambiguous with “*behavior*, *custom*, …”[27], respectively – with both nouns being in ablative case. It is worth noticing that appearance of the expression in our list also indicates potential issues in the boilerplate removal and filtration procedure applied during corpus processing.

In general, however, the beginning of the list looks “processable”.

4.1. At First Sight

The subsequent analysis is based on the first 400 lines of the candidate list. In the first step expressions in languages other than Latin (mostly English, but also French, German, Spanish and Italian) and also those representing proper names (mobile phones, apps, car and camera brand names, etc.) were manually sorted out.

As the list contained case information of the respective phrases, we were usually able to distinguish between the proprial use of a certain phrase and its use in its original meaning (alma mater vs. *Alma mater* / *Alma Mater* / *ALMA MATER*).

Such distinctions (verified by corpus queries) helped us to sort out expressions representing for example names of journals (*Ex Libris*), computer games (*Natus Vincere*), theatres (*Et Cetera*), music bands (*Status Quo*), etc.

In cases where the use of capital letters was not relevant, the respective frequencies could be aggregated. In our list, this was the case *De facto* (353), and *de facto* (2824), but not *De Facto* (601); *et cetera* (762), and *Et cetera* (328), but not *Et Cetera* (1406); *Alma mater* (964), and *alma mater* (2250), – here even *Alma Mater* (1141) was used in its original meaning in some contexts.

The results are presented in tables clustering the Latin expressions into lexical groups, and the respective tables are sorted according to the rank of the first item in the table. For each expression, the absolute and relative frequency (in instances per million) is also shown. The use of lowercase and uppercase letters is preserved.

4.2. Terms, Nomenclature, etc.

The most frequent expressions can be roughly divided into four groups: medicine terms, biological nomenclature, legal terms and expressions from religious discourse. Understandably, the respective expressions need not be used strictly in their terminological meaning in all contexts – a more precise analysis, however, would require much more time than was available for this work.

The resulting lists are shown in Tables 6 to 9 – the ten most frequent items are shown for longer lists.

Table 6. Medical terms

Rank	Freq	i.p.m.	Expression
2	15881	0.8030	*in vitro*
4	7671	0.3879	*in vivo*
48	1252	0.0633	*per os*
103	664	0.0336	*In vitro*
152	498	0.0252	*ex vivo*
159	471	0.0238	*spina bifida*

Table 7. Biological nomenclature

Rank	Freq	i.p.m.	Expression
3	8558	0.4327	*Homo sapiens*
9	4782	0.2418	*homo sapiens*
22	2692	0.1361	*Homo Sapiens*
28	2082	0.1053	*Candida albicans*
30	1928	0.0975	*Homo erectus*
56	1122	0.0567	*Homo habilis*
85	852	0.0431	*Aloe Vera*
245	337	0.0170	*Aedes aegypti*
283	304	0.0154	*Apis mellifera*
326	272	0.0138	*Candida Albicans*

Table 8. Legal/political terms

Rank	Freq	i.p.m.	Expression
19	2824	0.1428	*de facto*
23	2560	0.1294	*status quo*
73	973	0.0492	*de jure*
74	971	0.0491	*pro bono*
100	682	0.0345	*pro et contra*
106	660	0.0334	*post factum*
161	470	0.0238	*casus belli*
164	460	0.0233	*ex officio*
219	360	0.0182	*res publica*
227	353	0.0178	*De facto*

Table 9. Biblical/religious expressions

Rank	Freq	i.p.m.	Expression
37	1444	0.0730	*Opus Dei*
176	437	0.0221	*memento mori*
272	313	0.0158	*Memento mori*
289	300	0.0152	*ex nihilo*
377	239	0.0121	*Sola Scriptura*
394	234	0.0118	*Corpus Christi*

4.3. And All the Rest

The bulk of the remaining Latin expressions can be approximately classified as the Grudeva's Group I (from Table 1), though we admit that such categorization of some of the items may be disputable. As we only considered expressions consisting of at least two words, her item ranked 1 (etc) is naturally absent from our list, though its full form (et cetera, Et cetera) is still present there. Two other expressions from Group I (persona non grata and post factum) have been placed into our legal list, while the most frequent occurrences of nota bene (i.e. Nota Bene) have been sorted out as being mostly used as proper names.

Table 10 shows the twenty most frequent expressions from this group.

Table 10. Unclassified expressions

Rank	Freq	i.p.m.	Expression
13	3819	0.1931	*in situ*
20	2756	0.1393	*ad hoc*
26	2250	0.1138	*alma mater*
41	1408	0.0712	*terra incognita*
45	1282	0.0648	*alter ego*
63	1004	0.0508	*de novo*
66	997	0.0504	*tabula rasa*
77	964	0.0487	*Alma mater*
78	961	0.0486	*per se*
86	835	0.0422	*sui generis*
92	762	0.0385	*et cetera*
94	752	0.0380	*mutatis mutandis*
137	536	0.0271	*honoris causa*
140	529	0.0267	*modus vivendi*
143	525	0.0265	*modus operandi*
171	447	0.0226	*perpetuum mobile*
203	385	0.0195	*prima facie*
218	361	0.0183	*inter alia*
251	328	0.0166	*Et cetera*
252	327	0.0165	*urbi et orbi*

4.4. This Is the Beginning Only

Our present work was targeted more to getting an idea of what can be expected during a more profound analysis, than to receiving a "definite" classification of Latinisms in Russian texts.

The (semi-)automatically produced list exhibits some issues that could be most likely – at least partially – also tackled by automated procedures, e.g. by trying to get rid of "obviously" non-Latin expressions. But even if we managed to shrink the list size by the estimated 75%, it would still remain a lot of material to study.

5. Conclusions and Further Work

From the computational perspective, the presented work can be treated as a "proof of feasibility" of such a data-driven approach identification of foreign-language text fragments in a Russian corpus. We can say that it was successful, and it might be potentially used for searching not only of Latin phrases but of those of other languages as well. In the case of Russian, three other languages might be good candidates for similar research – English, French and German. In such a case, however, it would be reasonable to include one more step into the procedure: identification of all-Latin tokens (allowing also for accented letters to include French and German words). Such an arrangement might also decrease the total processing time, as the semi-product could be reused for several languages.

At the time of writing this paper, we were able to analyze and classify only a very small part of the candidate list produced. We not only want to process more of it but also provide the data to other researchers interested in Russian lexicology and lexicography.

A perspective area of research could also be a more systematic attempt of identification of Latin phrases written in Cyrillic script, such as "терра инкогнита", e.g. by applying the Latin tagger to Russian texts transliterated to Latin script. Though we can see some potential pitfalls here (such as inconsistencies in the transliteration), we believe that it is (at least) worth trying.

The described methodology could be also used to identify foreign-language text fragments in corpora of other languages. Our pilot experiment with an English corpus, however, indicated that it might be not that easy – the English morphological dictionary present in the TreeTagges's English language model [28] seems to include a great amount of Latin lexical items, thus making the simplistic approach of looking for "<unknown>" tokens problematic. We can, however, treat it as a next challenge.

This work has been, in part, funded by the VEGA Grant Agency, Project Nos. 2/0017/17 and 2/0103/19, respectively.

References

[1] Educational project "Polka". URL: https://polka.academy/articles/501 (accessed 19.01.2020).

[2] Jakubíček M., Kilgarriff A., Kovář V., Rychlý P., Suchomel, V. The TenTen Corpus Family // 7th International Corpus Linguistics Conference, Lancaster, July 2013. P. 125-127.

[3] Belikov V., Selegey V., Sharoff S. Preliminary considerations towards developing the General Internet Corpus of Russian // Komp'juternaja lingvistika i intellectual'nye tehnologii: Trudy mezhdunarodnoj konferentsii «Dialog–2012» [Computational Linguistics and Intellectual Technologies. Proceedings of International Conference «Dialog–2012»]. Moscow, RGGU. 2012. P. 37–49.

[4] Belikov V., Kopylov, N., Piperski, A., Selegey, V., Sharoff S.: Corpus as language: from scalability to register variation, [Korpus kak yazyk: ot masshtabiruyemosti k differentsial'noy polnote], Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialog 2013" [Komp'juternaja lingvistika i intellektual'nye tekhnologii: po materialam ezhegodnoy mezhdunarodnoj konferentsii «Dialog 2013»], vol. 12 (19), Moscow, RGGU. 2013. P. 84–95.

[5] Benko V. and Zakharov, V.P. Very large Russian corpora: New opportunities and new challenges // Kompjuternaja lingvistika i intellektuaľnyje technologii. 2016. P. 79–93.

[6] Shavrina, T.O. Differential Approach to Web Corpus Construction // Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference “Dialogue 2018” Moscow, May 30 – June 2, 2018.

[7] Benko V. Shavrina, T.O. Omnia Russica: Even larger Russian corpus // Proceedings of the International Conference “Corpus linguistics-2019”, June 24–28, 2019, St. Petersburg. Saint Petersburg University Press. 2019. P. 94–102.

[8] Comenius University in Bratislava UNESCO Chair in Plurilingual and Multicultural Communication. URL: http://unesco.uniba.sk/guest/ (accessed 19.01.2020).

[9] Benko V. Aranea: Yet Another Family of (Comparable) Web Corpora // Petr Sojka, Aleš Horák, Ivan Kopeček and Karel Pala (Eds.): Text, Speech and Dialogue. 17th International Confer-ence, TSD 2014, Brno, Czech Republic, September 8–12, 2014. Proceedings. LNCS 8655. Springer In-ternational Publishing Switzerland. 2014. P. 257– 264.

[10] The free dictionary: Winged. URL: https://idioms.thefreedictionary.com/winged (accessed 19.01.2020).

[11] Melnik N.V., Nuraleeva, G.I. Strategii tolkovaniy latinskikh krylatykh fraz v lingvo-personologicheskom aspekte [Strategies of Latin Phrases Explications in the Aspect of Linguistic Per-sonology] // Vestnik Kemerosvkogo gosudarstvennogo universiteta. 2014. № 3(59). P. 170-173.

[12] Petrova M.V. Sovar’ krylatykh vyraheniy. Moskva: Ripol Klassik. 2011. 280 p.

[13] Grudeva E.V. Kodovye pereklyucheniya s russkogo yazyka na latyn’ v russkikh tekstakh raznogo tipa (na materiale Natsinoal’nogo korpusa russkogo yazyka) [Code switching from Russian language to Latin in Russian texts of various types] In Vestnik Cherepovskogo gosudarstvennogo universiteta. 2014. №4 (57). P. 93–97.

[14] Grudeva E.V., Pavlova T. Funktsiinirovanie latinskikh vyrazheniy v russkikh tekstakh tekstakh raznogo tipa (na materiale Na-tsionlal’nogo korpusa russkogo yazyka) [Functioning of Latin Phrases in Russian Texts of Different Types (Based On the Material of National Corpus of the Russian Language)]. V mire nauki i isskustva: voprosy filologii i kul’turologii: Sbitnik statej po materialam XXXII mezhdunarodnoy nauchno-prakti-cheskoy konferentsii. Novivibirsk: SibAK. 2014. № 1(32).

[15] Natsional’nyj korpus russkogo yazyka: 2006–2008: Novye rezul’taty i perspektivy. St. Petersburg: Nes-tor-Istoriya. 2009. URL: http://ruscorpora.ru/corpora-biblio-2008.html (Accessed 19.01.2020).

[16] Grudeva E.V. Osobennosti funktsionirovaniya latinskikh krylatykh vyrazheniy s prityahatelnymi mestoimeniyami v russkikh tekstakh raznogo tipa (na materiale Natsionlal’nogo korpusa russkogo yazyka) [Functional Peculiarities of Latin Eloquent Expressions with Possessive Pronouns in Russian Texts of Different Types (as exemplified in Russian National Corpus)]. Universum: Filologiya i iskusstvovedenie. Electronic scientific journal. 2015. № 1 (15). URL: http://7universum.com/ru/philology/archive/item/1875 (Accessed: 19.01.2020).

[17] For a similar research in English it would be reasonable to consider also two non-alphabetic characters – an apostrophe and a hyphen. URL: https://nlp.fi.muni.cz/trac/noske (accessed 19.01.2020).

[18] Rychlý P. Manatee/Bonito A Modular Corpus Manager // 1st Workshop on Recent Advances in Slavonic Natural Language Processing. Brno: Masaryk University. 2007. P. 65–70. ISBN 978-80-210-4471-5.

[19] McGillivray B. Methods in Latin Computational Linguistics. Leiden/Boston: Brill. ISBN: 2211-4904. 2014.

[20] Schmid H.: Probabilistic Part-of-Speech Tagging Using Decision Trees. Proceedings of International Conference on New Methods in Language Processing, Manchester, UK. 1994.

[21] Language model. URL: https://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/data/Latin-parameter-file-readme (accessed 19.01.2020).

[22] Straka M., Hajič J., Straková J. UDPipe Trainable Pipeline for Processing CoNLL-U Files Performing Tokenization, Morphological Analysis, POS Tagging and Parsing. In Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016), Portorož, Slovenia, May 2016.

[23] Language model. URL: https://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/data/latin.par.gz (accessed 19.01.2020).

[24] Universal Dependencies. URL: https://universaldependencies.org/ (accessed 19.01.2020).

[25] Polit.ru: Civil society and civil policy. URL: http://polit.ru/article/2005/06/01/auzan/ (accessed 15.03.2020).

[26] Latdict. URL: http://www.latin-dictionary.net/ (accessed 13.01.2020).

[27] TreeTagges's English language model. URL: https://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/data/english.par.gz (accessed 19.02.2020).