A General Method Applicable to the Search for Anglicisms in Russian Social Network Texts

Alena Fenogenova, Ilia Karpov
National Research University Higher School of Economics,
Research and Development Institute KVANT
Moscow, Russia
{alenush93, karpovilia}@gmail.com

Viktor Kazorin
Research and Development Institute KVANT
Moscow, Russia
zhelyazik@mail.ru

***Abstract*—In the process of globalization, the number of English words in other languages has rapidly increased. In automatic speech recognition systems, spell-checking, tagging, and other software in the field of natural language processing, loan words are not easily recognized and should be evaluated separately. In this paper we present a corpora-based approach to the automatic detection of anglicisms in Russian social network texts. Proposed method is based on the idea of simultaneous scripting, phonetics, and semantics similarity of the original Latin word and its Cyrillic analogue. We used a set of transliteration, phonetic transcribing, and morphological analysis methods to find possible hypotheses and distributional semantic models to filter them. Resulting list of borrowings, gathered from approximately 20 million LiveJournal texts, shows good intersection with manually collected dictionary. Proposed method is fully automated and can be applied to any domain-specific area.**

I. Introduction

As English is currently the dominant language of business, economics, science, technology, and other fields, it has become one of the main sources of lexical borrowings. The influence of English on other languages grows rapidly [8], resulting in an increasing use of anglicisms. The number of English words is growing in Russian language as well, which raises a problem of finding new words that are not yet presented in dictionaries. As anglicisms are written in texts mostly with Cyrillic symbols, the task of automatic detection of such lexical items becomes a challenge.

The purpose of the present work is to propose an algorithm of automatic detection of new words in Russian texts borrowed from English. The phenomenon of anglicisms causes a range of thought-provoking questions for theoretical research; moreover, it is of great interest to practical applications, for instance, in spell-checking and morphological analysis. Knowledge of the way new words are generated would help creating a more accurate automatic natural language processing systems. The automatic detection of Cyrillic-written anglicisms in Russian text is a new, non-trivial and actual problem, in which it is also important to take into consideration the orthographic variation of English borrowings. To the best of our knowledge, there is a dearth of Russian-specific studies on this topic.

One of the most controversial issues is defining the notion of anglicism; it is a subjective lexicological question. In Russian texts, several types of English borrowings are usually presented:

- pure anglicisms (ex.: *ipad – айпад, fashion – фэшн, youtube – ютуб, freelance – фриланс, hardcore – хардкор, sorry – сорри,* etc.) – the word is written in Russian as it sounds in English;
- English roots, combined with Russian affixes (ex.: *gif +ка => гифка, flash+ка => флешка, creative+щик => креативщик, like+нуть => лайкнуть, twitt+нуть => твитнуть, forum+ок => форумок,* etc.) – the word has English root and some Russian flexion;
- abbreviations (ex.: *CNN – сиэнэн, IBM – айбиэм, ZIP –зип* etc.)
- composites, containing multiple English words (ex.: *life +hack – лайфхак, old+school – олдскул* etc.)

The variety of anglicisms types is not limited to this list. Some of the words were so quickly introduced into the Russian lexicon that they are not perceived as new words anymore.

The criteria of defining the word as anglicism need to be considered as a set, because entrance velocity depends on the word's popularity among the native Russian language speakers. For example, words like *бизнес* [*business*], *маркетинг* [*marketing*], *пирсинг* [*piercing*] hardly can be classified as new borrowings, but domain-specific words like *сервер* [*server*], *прокси* [*proxy*] can be, though they have appeared in Russian relatively long ago. It is also important to distinguish between the loan words and the word-formation derivatives. For example, substantive *футболка* [*Tshirt*] or verb *гамать* [*to game*] cannot be defined as anglicisms, as they have acquired semantics that was initially assumed as shown in this example: *football => футбол => футбол+ка*}.

Another difficult case, which should be taken into account, is a sense ambiguity and the problem of disambiguation. For example, the word *пост* isn't anglicism within the meaning of fasting, but can be considered as borrowing not only in the meaning of position, held by someone (*city mayor's post*), but also in the meaning of the online text message (*forum post*).

ISBN 978-952-68397-8-3 (PDF)

We propose a set of informal criteria to define new anglicisms. From derivation perspective, if the word has a wide range of derivatives, we assume it is used in language for an extended time (ex.: *пенал – пенальчик [penal – penal'chik]* but *ноутбук* [*notebook*]. Spelling is another criterion – if the norms of orthography are not well-established, the word is likely to be new (ex.: *флешка – флэшка* (from *flesh*)). Grammar and phonetics mark anglicisms as well – the rules of grammar can be broken on borrowed words (for instance, English borrowed adjectives are not inflected) or for example, the *e* tones down when the borrowing is long-standing (д`**е***таль* [*d'etal'*] but *мод***е***ратор* [*moderator*]). The loan words can be checked in corpora by frequency: if the word appeared not so long ago and the percent of its occurrences is increasing in recent years – it is a new borrowing. All these features are criteria of anglicisms' definition and, at the same time, they are good clue for their detection. For the practical application of the proposed method it is important that a Russian word can be automatically linked to its English cognate.

In this paper we present a complex approach for the automatic detection of anglicisms included in Russian texts. Our algorithm does not contain any prepared, manually acquired data; instead it copes with the new texts and reckons for possible new borrowings. We carry out the comparison between the anglicisms our approach can handle with and anglicisms of manually collected dictionary. Our elaboration can improve systems of automatic speech recognition, spell-checking, tagging and other tasks in the field of natural language processing as well as it can be of great interest to the linguistic community, studying language contact processes.

II. Related work

The phenomenon of anglicisms is occurring in different languages, notably in European, for instance, Italian [5], French [17], Croatian [7] and many others. A various types of lexical borrowings in European languages and precise researches of their causes are described in detail in the work [18]. Types of borrowings and language contacts vary significantly depending on a particular language. In the work [1] author circumstantially describes these types and presented the English inclusion classifier based on the German data.

One of the strongest factors of interlingual influence is a genetic similarity. A great number of researches are done in this field on the basis of Germanic languages, such as Norwegian [11], Danish [9] German [12]. In the latter work authors proposed methods for automatic detection of anglicisms and applied them to Afrikaans and German languages. They developed a set of features and collected and annotated a German IT corpus to evaluate them. Set is consists of the following features: grapheme perplexity, G2P confidence, English Hunspell lookup, Wiktionary lookup, and Google hits count. None of these single features rely on annotated text with anglicisms for training. Combining features authors reach 75.44 % F-score.

Another interesting research about automatic detection of English neologisms was based on Norwegian language [3]. In this paper author has tested rule-based (manually constructed regular expressions), lexicon-based (lexicon lookup methodology), chargram-based (list of chargrams) and combinatory methods for retrieval of anglicisms in Norwegian texts. The experiments in both works have shown that the optimal results are gained when the combination of methods is applied. However, for Russian language such techniques are not appropriate due to the various factors. For example, even the lexicon-based feature can not be straightforward used, as the anglicisms are written in Cyrillic symbols and their entries can not be directly checked in wordlists, Hunspell or Wiktionary lookup.

In Statistical Machine Translation the problem of out-of-vocabulary (OOV) words are ubiquitous and actual. The current trend in SMT is to use deep learning and neural networks techniques [13],[10], which achieve promising results. In paper [24] researchers on a Chinese material try not to find the direct translation for the unknown words, but determine the semantic function of such words and keep the semantic function unchanged in the translation process. From this perspective the machine translation task has a lot in common with the problem of detection of anglicisms, as in both cases we need to determine the semantic function of the specified words using the context. It's not surprising that some of the works in machine translation [23] employs the distributional semantic models. In the work [21] author tries to create a lexical borrowing model, demonstrated it in machine translation. The translation candidates are produced by phonetic and semantic (word2vec method) features.

A significant amount of theoretical works about anglicisms in Russian language are written [4],[6],[22],[20] and even the manually developed dictionary of anglicisms is available online (http://anglicismdictionary.dishman.ru/slovar). In the work [16] study of neologisms and loan words frequently occurring in Facebook user posts were presented. The authors collected a dataset of about 573 million posts written during 2006–2013 by Russian-speaking Facebook users half-automated. As a result they produced a list of 168 neologisms, including anglicisms and attempted to make etymological classification and distinguished thematic areas of these neologisms. However, in Russian academic circles the problem of automatic detection of anglicisms is still out of the scope.

III. Methodology

The general method is based on the idea of simultaneous scripting, phonetics and semantics similarity of the original Latin word and its Cyrillic analogue. We assumed words, that sound or script in the same way from one side and close in word2vec model from another side most likely to be language borrowings. As shown in figure 1, having 2 corpora, (Russian, English) we take all frequently (more than 30 in this work) words and generate a list of hypotheses for each pair of words. We make a list of possible transcriptions and transliterations for English word and compare them with Russian normal form and preformed root by Levenshtein distance. We get the Levenshtein distance threshold as a function of words length,

but the maximum threshold is set to 4 for normal forms and 3 for roots to avoid combinatorial explosion. We also reduce the hypotheses amount by means of English-Russian dictionaries – if some word's translation is close to the original we don't need to check it in any other way, though it gives us only "well-known" anglicisms.

Having a list of hypotheses, we consistently check them by two distributive semantic models. If the word is used by Russian speakers simultaneous in Cyrillic and Latin (ex.: *ноутбук – notebook*) both scripting variants will be used in the same context, that can be proved by SkipGram model, trained on Russian corpora. If the word is very close to it's English analogue, but we failed to found it in SkipGram top, we translate it's left and right context and use CBOW model, trained on English corpora to find out if the context is close to English's analogue contexts. The second method doesn't shows the borrowing direction, we cannot be sure if it is the English word, used by Russian speakers or conversely Russian word used by English speakers, but it looks trustworthy, that Russian speakers are highly influenced by English, given at schools, domain-specific literature and Web-resources, so the most borrowing came from English to Russian.

The rest of this section is divided into two blocs, describing all inner steps of hypotheses generation and hypotheses filtering.

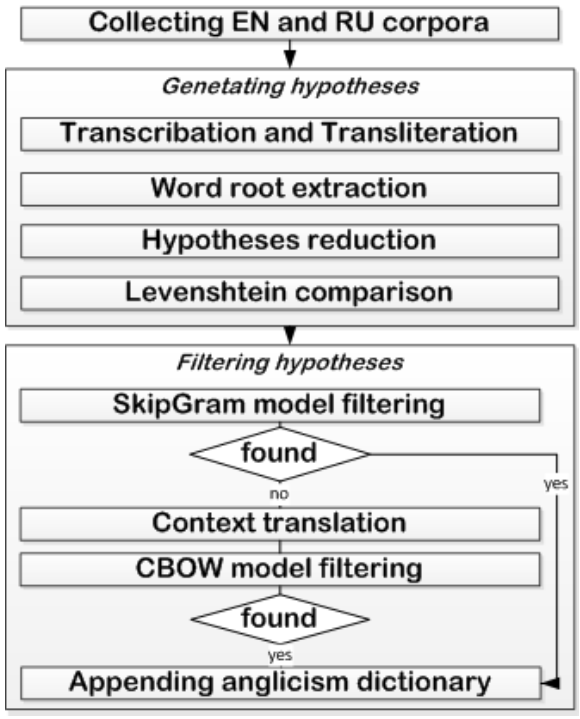

Fig. 1. General method description

A. Hypotheses generation

The proposed approach is based on the fact, that language speakers tend to save phonetic and orthographic properties of the borrowed word. We assume that borrowed word was either transliterated or transcribed from English to Russian.

In case of transliteration, speaker is supposed to convert text from Latin script to Cyrillic (Cyrillization) using internal intuition about the writing system. We have not found any official standard for Cyrillization of Latin letters, but there is a list of contradictory standards to transliterate the Russian language from the Cyrillic script into the Latin alphabet (Romanization) such as ICAO, GOST 7.79 – 2000, 16876 – 71, 52535.1 – 2006b and ISO/R 9 (http://en.wikipedia.org/wiki/Romanization_of_Russian). Reversing these rules we gained context-free generative grammar that converted Latin word to multiple Cyrillic scripting hypotheses.

In case of transcribing, speaker is supposed to save word's phonation while writing English word with Cyrillic script. Considering that English phonetics differs from speaker to speaker and contains lots of exceptions, we used only in-vocabulary lexis with pre-defined transcriptions, gained from joint Cambridge Advanced Learner's Dictionary, Cambridge Academic Content Dictionary and Cambridge Business English Dictionary (http://dictionary.cambridge.org/dictionary/english). We took both English and American transcriptions and developed a context-dependent grammar, based on practical transcription of English named entities, proposed by Gilyarevsky [2].

In both cases we supposed that the speaker is usually trying to make visual presentation as close as possible, so we added additional rules, that convert similar looking characters even if they conflict with existing grammar *A=>A*, *E=>E*, *Y=>Y*, etc.).

One of the most frequently used way of anglicism generation assumes that Russian word contains English word as root with Russian affixes added, so we developed the neural network based root extraction method, trained on 97,000 normal form –root pairs, extracted from WikiDictionary. This task supposes that the neural network generates new sequence of characters from the existing one, provided that the resulting sequence is a substring of the input sequence so we used Recurrent neural network work (RNN), that showed good results on similar tasks [18]. LSTM-based bidirectional network was used as an input layer and dense layer with softmax activation function as output layer. For train network we used ~500,000 word forms set gained from the initial 97,000 normal forms and corresponding roots. Each character was transformed to one-hot-encoding vector. Keras framework was used for RNN training. RNN was used only for non-dictionary words of corpora. Resulting list contained 142,152 unique words and their roots, encountered more than 30 times in the Russian corpora. We haven't done proper evaluation of this method, but it seems that trained RNN sometimes combines root with prefix, but generally works well with one-root words. It totally fails to extract multiple roots, that can cause some errors in the proposed method and needs to be improved in future work.

Additional hypotheses were generated to predict composite anglicisms, consisting of two English words. First, we've generated a list of bigrams, based on word collocations weighting function, proposed by Mikolov [14]. However, even with part of speech restrictions on the amount of bigrams

hypotheses, it does not help us to avoid the generation of a huge set of anglicisms bigrams, most of which are not composites at all. To reduce production costs we used another approach – we assumed that each Russian word is a compound word and separated it on two possible parts. For this purpose we used the implementation of the maximum matching algorithm (HashTag Splitter - https://github.com/matchado/HashTagSplitter) used in Natural Language Processing to split compound words or hashtags to multiple words.

Next step in hypotheses generation was to select appropriate patterns by Levenshtein distance. As was described above we have got the sets of roots for Russian words and the Cambridge-based set of possible English transcriptions and transliterated patterns. We compared all possible combinations of lemma-transcription, lemma-transliteration, roots-transcription, roots-transliteration for each Russian word. Edit distance was modified by assigning special weighted penalties for some cases. For example, we do not penalize for edits with spaces and hyphens, while edits *э* and *е* are received weight -0.5 instead of -1 (compare: *флеш* vs. *флэш*, or *джазфанк* vs. *джаз-фанк* vs. *джаз фанк*). For producing the final hypotheses the threshold of edit distance was selected empirically and was equal to 2 for roots with length more than 3 and threshold 1 for short roots. For lemmas respectively, if the length was more than 5 – edit distance's threshold was 3, if less – 2. Resulting hypotheses examples are shown at the Table I.

TABLE I. HYPOTHESES GENERATION EXAMPLES

EN word	EN-RU	TR-RU	RU-EN	RU word
football	футбол(0)	футбол(0)	footbol(2)	футбол
brainstorm	браинсторм (1.5)	брэйнстом (1)	bryeynshtorm (3)	брейншторм
fashion	фашион(2)	фэшэн(1)	feshn(3)	фэшн

In Table I EN word – original English word, found in English LiveJournal, RU word – original Russian word, found in Russian LiveJournal, EN–RU and RU–EN – transliteration result, TR–RU – transcription result, numbers in brackets are corresponding Levenshtein distances (LD) between the hypothesis and the original word.

B. Hypotheses validation

We propose two methods to filter hypotheses, obtained at the previous step. First one is based on the fact that many anglicisms are written by users in both English and Russian spelling in social network texts. Algorithm 1 describes Levenshtein distance dependent filtering method.

Lets denote hypotheses set as H, anglicisms set as A. Any $h \in H$ consist of *h.rus* – candidate to anglicism, *h.eng* – prototype for anglicism, *h.editDist* – Levenshtein edit distance between *h.rus* and *h.eng*. In set A we will keep pairs *(h.rus, h.eng)* if *h.rus* – anglicism.

Algorithm 1 Hypotheses validation

```
1: topByDist = {1000, 100, 10}
2: A = Ø
3: for all h ∈ H do
4: nearestVecs = w2vModel.getMostSimilar(h.rus)
5:     if h.eng in top topByDist[h.editDist] nearestVec then
6: A.add((h.rus,h.eng))
7:   end if
8: end for
```

Many anglicisms are rarely used by Russian speakers in original (English) spelling, that makes many relevant hypotheses to be lost while using SkipGram filtering. We propose translation and context search with CBOW model, trained on English texts. For each hypotheses all contexts, containing 5 words left and 5 words right the hypotheses are translated. If top 100 most relevant English words for each context contains English analogue of the hypotheses in more than 50% cases, we consider the hypotheses to be proved.

IV. EVALUATION

This section describes method evaluation using 10 million Russian and 10 million English texts from LiveJournal blog platform. LiveJournal provides representational lexical corpora, that covers a large variety of themes, written by users of different age, interests and places of living in one place. Using single blog platform keeps us safe from combining texts from different resources with their specific lexis, markup and conversation style, that can cause negative effects of distributional semantics methods. Besides, LiveJournal, as distinct from other social networks, such as Facebook, Twitter and VKontakte, has less proportion of plagiarism, so we do not have to filter duplicates and advertisements.

A. Data collection

We collected the list of 100,000 Russian and English authors, available at top bloggers LiveJournal rank and found mode authors from comments, given to their posts. We made a randomly choose authors from approximately 8 million list and downloaded their posts with comments, starting from 2010 until we got 10 million texts for each language. We cleaned up the texts from html-markup kept by API, made language detection, graphematical and morphological parse. NTextCat (https://github.com/ivanakcheurov/ntextcat) library, trained on n-grams from Wikipedia, was used to predict language from text, Mystem (https://tech.yandex.ru/mystem/) was used for Russian texts and Stanford CoreNLP [14] engine for English.

We had to collect more texts after the parsing was done as nearly 20% of materials contained no text or texts, shorter than 200 characters, and nearly 8% of materials were classified as other languages.

B. Testing set

For result's evaluation we need to distinguish which words are to be concerned as anglicisms and which are not. For this

purpose the list of anglicisms that are already proved to be valid need to be collected. The list consists of filtered words from two resources: the dictionary of teenager's slang (http://teenslang.su) and A.I.Dyakov dictionary of anglicisms.

The dictionary of anglicisms by A.I.Dyakov is the basis of our list. In 2010 "The Dictionary of Anglicisms of the Russian Language" was edited in Novossibirsk. The electronic version of the Dyakov dictionary is available online since 2014 and it contains about 20,000 lexical items, among them 1000 collocations. The dictionary offers borrowings from a wide range of living spheres: economics, IT, marketing, etc. It also involves loan words from spoken language, various slangs, professional jargons and profanities. It is the most detailed and comprehensive anglicism's dictionary available for Russian language. The author points out the deep inter-integration of two languages and distinguishes in his dictionary both the pure borrowings and Russian word-formation derivatives on the basis of anglicisms. For example, the cases like substantives with Russian suffixes *-ция* and *-ость*, which are duplicated form English words on *-tion*, are quite thought-provoking. Adjectives are hard to define if they are derivatives or borrowings as well (ex: *радиоактивный* [*radioactive*], *чековый* [*checking*], *янговый* [derivation from *young*], *тюнинговый* [derivation from *tuning*]). A.I.Dyakov notes that the choice of inclusion the word in the anglicism`s list is subjective and we share this point of view. Our filtered list of anglicisms do not include described cases of adjectives and substantives.

Except for the Dyakov dictionary of anglicisms we added in final sample the teenager's slang dictionaries entries. The dictionary of teenager's slang is online resource which is constantly refilling by common users, that are used in their active lexicon new words. The words before inclusion in our list were preliminary filtered by mark "from eng". In total, filtered list of anglicisms from Dyakov and teenager's slang (DTS) dictionary contains about 16,000 lexical items.

C. Experiments

Comparison of Levenshtein distance values against the size of the resulting vocabulary is provided in Table II. Columns Dictionary, SkipGram and CBOW correspond to 3 filtering strategies described above. CBOW model search was only for 380 words with LD < 1, due to limited translation resources.

Result column shows total amount of words, found by all strategies. We found only 4,321 words of the joint dictionary in our corpora, DTS column shows intersection between our results and these words.

TABLE II. COMPARISON WITH MANUALLY COLLECTED DICTIONARY

	Dictionary	SkipGram	CBOW	Result	DTS
LD≤1	430	1552	380	1362	863
LD≤2	980	1061	380*	2421	1207
LD≤4	1021	36480	380*	37881	1289

Table II proves that higher Levenshtein distance produces more hypotheses, found in distributive models but dramatically increases error of the second kind. Resulting intersection demonstrates the demand of more textual data, however, the addition of it was complicated by the absence of the complete anglicisms list for this task, so we expanded DTS dictionary as follows:

We have manually annotated all hypotheses with LD < 1, missed at the DTS dictionary to evaluate *Precision, Recall* and *F1-scores*. Evaluation showed 283 valid anglicisms against 222 wrong words. Assuming that all words found by our method and contained in the DTS dictionary are true anglicisms, we obtain predicted anglicisms, found in the test set, *True Positive* = 283 + 863 = 1146, and *False Positive* = 222. Assuming that we have found only 283 + 863 of 4321 + 283 pairs in the testing set, we obtained *False Negative* = 3458 and *True Negative* = 0. Results are shown in Table III.

Proposed metric is very approximate since there are only 2417 of 4321 pairs (both Russian word and English analogue) in the SkipGram model which means that only 0.56% of the words could be found in Russian corpora, and 1904 pairs could be found only by means of context translation.

TABLE III. CONFUSION MATRIX FOR ANGLICISMS DETECTION TASK

	Predicted Angl-s	Predicted non Angl-s
Observed Angl-s	1146	3458
Observed non Angl-s	222	0

According to these assumptions, we obtained *Precision* = 0.84, *Recall* = 0.24. Overall *F1-score* = 0.38.

V. DISCUSSION

Resulting method was able to find rather complex words like (*ретвитнуть – retweet)* and (*скетч – sketch*). It was also capable to match word phrases like (*киберспорт – cyber + sport*) and words missing in the manually collected dictionary like (*кикстартер – kickstarter*) and (*айпишник – ip*). CBOW model shows promising result in spite of computational complexity. For example, we failed to find (*скейтборд – skateboard*) in SkipGram model even though LD = 0 for this pair, but finally added it with CBOW model. There are still many errors in root extraction, that causes the system to find cases like (*декорирование – deco*) and there are distributive models errors like (*клип* [*clip*] *– creep*) as well, but most of the results looks trustworthy. Overall *F1-score* = 0.38 seems to be good result, given very rare and domain-specific anglicisms in the DTS dictionary.

The main problem of the proposed method is a slow hypotheses generation process due to combinatorial explosion and very expensive translation procedure.

Future steps should be done in the following directions:

- evaluate texts from other social networks like Twitter, as 10M LiveJournal corpora has shown poor intersection with manually collected dictionary;

- use character-based clustering algorithm instead of comparing all-possible combinations of original words and hypotheses with Levenshtein distance;
- root extraction function requires optimization to be suitable for multi-root words;
- morphological preprocessing should be modified as we took only the most probable variant of word's normalization, that may cause loss of some hypotheses;
- some transcription rules may be modified according to the specific of Russian language. In particular, if there is no such diphthongs in Russian language it's pronunciation may change from the original.

VI. Conclusion

The main aim of this article was to present general methodology to search foreign borrowings in Russian texts and their analogues in English. Proposed method was evaluated on LiveJournal platform corpora and compared with existing manually collected resources. About 1,150 of 4,300 words found by our method presented in the manually collected dictionaries, the rest of the words seem to be new anglicisms, that have not been systemized yet. Found anglicisms list is a great source for researches of language contact processes and can be used as an external dictionary to increase the quality of various applied tasks, such as morphological analysis, spell-checking, sentiment detection etc. Proposed method is fully automated and may be used in machine translation and corpus-based linguistics tasks as a linguistic assistant. All code, along with gathered dictionaries is available online at github (http://github.com/lab533/anglicisms).

Acknowledgment

The reported study is supported by RFBR grant №16-29-09614 "Principles and technologies of machine processing for social network texts in Russian" and Russian Academic Excellence Project '5-100'.

We thank Svetlana Timoshenko, Tatiana Krylova, Eduard Klyshinskiy for fruitful discussions and valuable comments on proposed work.

References

[1] Beatrice Alex, "Automatic detection of English inclusions in mixed-lingual data with an application to parsing", *PhD thesis.* University of Edinburg, 2007.

[2] Amineva et al, *Practical transcription of Noun groups, 2-nd edition.* 2006.

[3] Gisle Andersen,"Assessing algorithms for automatic extraction of Anglicisms in Norwegian texts", *Corpus Linguistics 2005.*, 2005.

[4] M.A. Breiter."Anglicisms in Russian language: history and perspectives.", *Vladiwostok, Dialo-MSU.,* 1997.

[5] M. Cristina, Caimotto and Alessandra Molino, "Anglicisms in Italian as alerts to greenwashing: a case study", *Critical Approaches to Discourse Analysis across Disciplines.,* vol. 5.1, 2011, pp. 1–16.

[6] A.I. Dyakov."The causes of borrowings of anglicisms in modern Russian language", *Yasyk i Cultura.,* vol. 1, 2003.

[7] Blavzenka, Filipan-Zigni'c. *Neologisms in modern Croatian Language. Philology, technology and sociology related researches at Eotvos Jozsef College.* Eotvos Jozsef College, 2008.

[8] Manfred Görlach et al, "A dictionary of European Anglicisms". *A Usage Dictionary of.,* 2001.

[9] Henrik Gottlieb, "The impact of English: Danish TV subtitles as mediators of Anglicisms", *Zeitschrift fur Anglistik und Amerikanistik.,* 1999.

[10] Caglar Gulcehre et al,"Pointing the Unknown Words", *preprint arXiv: 1603.08148.,* 2016.

[11] Marita Kristiansen, "Detecting specialised neologisms in researchers' blogs", *Bergen Language and Linguistics Studies.,* vol. 3.1, 2013.

[12] Sebastian Leidig, Tim Schlippe,and Tanja Schultz."Automatic Detection of Anglicisms for the Pronunciation Dictionary Generation: A Case Study on Our German IT Corpus", *Spoken Language Technologies for Under-Resourced Languages.* 2014.

[13] Minh-Thang Luong et al. "Addressing the rare word problem in neural machine translation", *arXivpreprint arXiv:1410.8206,* 2014.

[14] Christopher D. Manning et al. "The Stanford CoreNLP Natural Language Processing Toolkit", *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics.,* 2014.

[15] Tomas Mikolov et al."Distributed Representations of Words and Phrases and their Compositionality", *Proceedings of NIPS.,* 2013.

[16] N.A. Muraviev, A.I. Panchenko, and S.A. Obiedkov. "Neologisms on facebook", *Dialog.,* 2014.

[17] Michael D Picone, *Anglicisms, neologisms and dynamic French.,* John Benjamins Publishing, vol. 18, 1996.

[18] Virginia Pulcini, Cristiano Furiassi,and F'elix Rodr'iguez Gonz'alez, "The lexical influence of English on European languages", *The Anglicization of European Lexis,* vol. 1., 2012.

[19] Ilya Sutskever, Oriol Vinyals, and Quoc V. Le,"Sequence to sequence learning with neural networks", *Advances in neural information processing systems.*, 2014, pp. 3104–3112.

[20] G.G. Timofeeva, *New english borrowings in Russian language: spelling, pronunciation.,* Yuna, 1995.

[21] Yulia Tsvetkov and Chris Dyer."Cross-lingualbridges with models of lexical borrowing", *JAIR55.,* 2016, pp. 63–93.

[22] E.F. Volodarskaya,"Borrowing as a reflection of Russian-British contacts", *Voprosy yasykosnania.,* vol.4, 2002, pp. 96–118.

[23] Yair Wolf et al. "Joint word2vec networks fo bilingual semantic representations", International Journal of Computational *Linguistics and Applications.,* 2014.

[24] Jia-Jun Zhang, Fei-Fei Zhai, and Cheng-Qing Zong, "A Substitution - Translation - Restoration Framework for Handling Unknown Words in Statistical Machine Translation", *Journal of Computer Science and Technology.,* vol. 28.5, 2013, pp. 907–918.