Computational Linguistics and Intellectual Technologies:
Proceedings of the International Conference "Dialogue 2017"

Moscow, May 31—June 3, 2017

COMPARATIVE ANALYSIS OF ANGLICISM DISTRIBUTION IN RUSSIAN SOCIAL NETWORK TEXTS

Fenogenova A. S. (alenush93@gmail.com),
Karpov I. A. (karpovilia@gmail.com),
Kazorin V. I. (zhelyazik@mail.ru),
Lebedev I. V. (innlebedev@gmail.com)

National Research University Higher School of Economics,
Research and Development Institute KVANT, Moscow, Russia

Due to the process of globalization, the number of English borrowings in different languages is constantly growing. In natural language processing (NLP) systems, such as spell-check, POS tags, etc. the analysis of loan words is not a trivial task and should be resolved separately. This article continues our previous work on the corpus-driven Anglicism detection by proposing an improved method to the search of loan words by means of contemporary machine translation methods. It then describes distribution of the borrowed lexicon in different online social networks (OSN) and blog platforms showing that the Anglicism search task strongly depends on corpus formation method. Our approach does not contain any pre-prepared, manually acquired data and gives a significant automation in Anglicism dictionary generation. We present an effective dictionary collection method that gives the same coverage compared to random user selection strategy on a 20 times smaller corpus. Our comparative study on LiveJournal, VKontakte, Habrahabr and Twitter shows that different social, gender, even age groups have the same proportion of Anglicisms in speech.

Key words: Anglicisms, distributive semantics, social media texts, semantics, vector representation

СРАВНИТЕЛЬНЫЙ АНАЛИЗ РАСПРЕДЕЛЕНИЯ АНГЛИЦИЗМОВ В РУССКИХ ТЕКСТАХ СОЦИАЛЬНЫХ МЕДИА

Феногенова А. С. (alenush93@gmail.com),
Карпов И. А. (karpovilia@gmail.com),
Казорин В. И. (zhelyazik@mail.ru),
Лебедев И. В. (innlebedev@gmail.com)

Национальный исследовательский университет
Высшая школа экономики, Научно-исследовательский
институт «Квант», Москва, Россия

В связи с процессом глобализации, наблюдается рост количества английских заимствований во многих языках мира. Поиск подобных заимствований представляет интерес как для теоретических исследований в области языковых контактов и межъязыкового взаимодействия, так и в прикладных задачах, например при разработке средств морфологического анализа, исправления опечаток и машинного перевода.

Данная работа продолжает выполненное авторами ранее исследование в области выявления англицизмов. В работе предложен улучшенный метод поиска английских заимствований в том числе с методами линейного отображения векторных пространств двух языков. Отличительной особенностью подхода является работа без подготовленных заранее словарей и собранных вручную коллекций.

Также рассматриваются вопросы распределения заимствований в различных корпусах русскоязычных пользователей социальных сетей. Предложена эффективная стратегия автоматического поиска текстовой информации, позволяющая уменьшить размер корпуса в 20 раз по сравнению со случайным сбором при сопоставимой полноте словаря. Сравнительный анализ материалов таких ресурсов как Живой Журнал, Вконтакте Твиттер показывает равномерность распределения заимствований в письменной речи пользователей различного пола и возраста.

Ключевые слова: поиск англицизмов, лексикография, дистрибутивная семантика, социально-сетевые тексты

1. Introduction

The widespread use of English in the process of globalization continues to have a tremendous impact on development of different languages, namely, the number of English words in them is growing rapidly. The phenomenon of Anglicisms is occurring in languages all over the world, and the Russian language is not an exception. In the field of natural language processing this tendency raises a problem, finding new words (loan words) that are not yet presented in dictionaries. The automatic detection of Cyrillic-written Anglicisms in Russian text is a new, non-trivial and actual problem, especially as it is representative of the texts of social networks. People commonly use loan words and orthographic variation of loan English borrowings in a significant way.

The notion of an Anglicism can be defined in various ways; what can be regarded as "true", as an Anglicism is a rather subjective issue. There are several types of English borrowings that we aim to detect:

- pure Anglicisms (*ex.: iPad—айпад, fashion—фэшн, YouTube—ютуб, etc.*)—the word written in Russian as it sounds in English;
- English roots, combined with Russian affixes (*ex.: gif+ка => гифка, от+football+ить => отфутболить, like+нуть => лайкнуть, etc.*)—the word has an English root and some Russian flexion;
- abbreviations (*ex.: LOL—лол, ZIP—зип, etc.*)
- composites (*ex.: life+hack—лайфхак, old+school—олдскул etc.*)—words with two English roots.

For the practical application of the proposed method it is important that a Russian word can be automatically linked to its English cognate.

A significant amount of theoretical works about integration of Anglicisms in Russian language, social-linguistic studies and interlanguage research are written (Chachibaia etc. 2005; Proshina, 2016; Janurik, 2010; Yaniv, 2016). The work (Chugunova etc. 2016) presents detailed classification of Anglicisms in Russian and continues the research of their adoption and origin. The authors (Muraviev, etc. 2014) study neologisms and loan words frequently occurring in Facebook user posts. The authors half-automatically collected a dataset of about 573 million posts from Russian-speaking users (written during the period from 2006 till 2013). As a result, authors produced a list of 168 neologisms, including Anglicisms and attempted to make etymological classification and distinguished thematic areas of these neologisms. Some research is devoted to classification of the modern Anglicisms on the Internet (Bylatcheva etc. 2016), others pay attention to the comparative studies of languages occupied by Anglicisms, as in the work made by Balakina (Balakina, 2011), where the author compares lexical items in Russian and German blogs. In one of his latest works (Dyakov, 2016) A. I. Dyakov classifies loan words and proposes an adaptation model of the Anglicism—the scheme of dynamic process in a Russian-speaker's thesaurus, frequency of use and mechanism of adoption. Over 10 years he manually collected more than 20,000 borrowed lexical items and from this considerable set of Anglicisms he created the Anglicism Dictionary1 available online.

For the Russian language, the method applicable to search for English loan words and their analogues in Russian social network texts was presented by authors of this paper (Fenogenova etc., 2016). The proposed general methodology on the material of LiveJournal texts was able to detect 1,146 Anglicisms based on 20 million LiveJournal texts and comments, but the proposed approach was limited by (a) the computational expense of machine translation procedure, proposed in the work, (b) the low fraction of Anglicisms in the collected corpus. Though the proposed method demonstrated relatively high recall, we failed to find many real-life Anglicisms due to their

[1] http://anglicismdictionary.dishman.ru/slovar

absence in our corpus. Thus, corpus formation (or network walk) strategy appears to be more of a vital problem rather than the hypotheses generation or the filtering strategy. Contributions to the present study are as follows:

- Modified Anglicism detection method: an approach to linear mapping between distributive vectors in different languages (Mikolov etc., 2013) was used instead of machine translation.
- Anglicism variety analysis: Anglicism distribution is independent to the data source and user age, sex, geographical location.
- Dictionary growth strategy: dictionary size is an asymptotic function of corpus size. The same amount of Anglicisms can be found on smaller corpus by means of effective data collection strategy.

2. Anglicism Detection Algorithm

The method is based on the idea that the original Latin word is similar to its Cyrillic analogue in scripting, phonetics and semantics. We assume that words are likely to be borrowed if they sound or script in the same way as their English analogues. At the same time loan words and their original equivalents should be close in the distributional semantics model. From the corpus of social network texts we take words, mentioned more than 30 times and generate a list of hypotheses for each pair of words. Next, we make a list of possible transcriptions and transliterations from English words and compare them with Russian tokens by Levenshtein Distance. We get the Levenshtein Distance threshold as a function of word length, but the maximum threshold is set to 3 for the normal forms. As a result we get a list of hypotheses pairs and check them by distributional semantics in the following ways. The general architecture of our method is presented in picture 1.

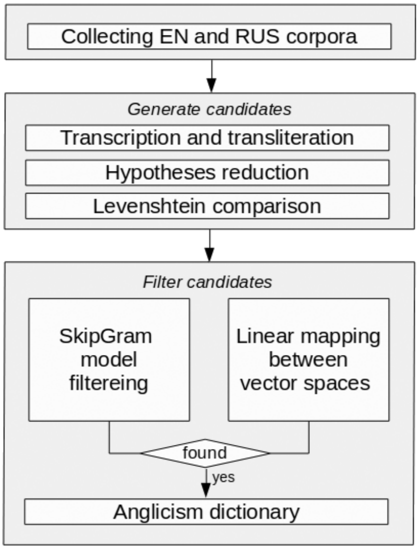
Collecting EN and RUS corpora
Generate candidates
Transcription and transliteration
Hypotheses reduction
Levenshtein comparison
Filter candidates
SkipGram model filtereing
Linear mapping between vector spaces
found
yes
Anglicism dictionary

Fig. 1. General algorithm

First, we verify our candidate from the hypotheses list appears in the model and has a Latin spelling as the most similar word that is equal to the English hypothesis candidate. Let us denote a hypotheses set as *H*, Anglicisms set as *A*. Any $h \in H$ consists of *h.rus*—a candidate to Anglicism, *h.eng*—prototype for Anglicism, *h.editDist*—Levenshtein edit distance between *h.rus* and *h.eng*. If *h.eng* in the top *n* nearest vectors, *h.rus* is proved to be Anglicism and we will form pairs (*h.rus*, *h.eng*) in set A.

Algorithm 1. Hypotheses validation

```
topByDist = {1000, 100, 10}
A=$\emptyset$
for all $h \in H$ do
    nearestVecs = w2vModel.getMostSimilar(h.rus)
  if h.eng in top topByDist[h.editDist] nearestVec then
    A.add((h.rus,h.eng))
 end if
end for
```

However, this method cannot cope with cases when the SkipGram model does not contain an English candidate. To solve the problem above, we trialed the method proposed by Mikolov (Mikolov etc., 2013) on our data. The English word2vec model was built and the linear mapping between vector space of English language and vector space of the Russian model was learnt. For linear mapping, we selected matrix *W* that minimizes

$$\sum_{i=1}^{m}\frac{1}{2}\|Wx_i - z_i\|^2,$$

where $x_i, z_i - i$—the pairing of an English word and its Russian translation. Next, mapping is provided between the linear vector that corresponded to the *hypothesis.rus,* and the vector of the word translation from the English vector space. The final step was to check the nearest top N vectors, if *hypothesis.eng* was proven to be in the list, *hypothesis.rus* was considered to be an Anglicism.

Algorithm 2. Hypotheses validation with translation

```
N = 100
A=$\emptyset$
for all $h \in H$ do
  vec = mapToRussianW2VSpace(h.eng)
  if h.rus in top N w2vModelRussian.nearestVec(vec) then
    A.add((h.rus, h.eng))
 end if
end for
```

3. Comparative experiments

For this study four datasets (LiveJournal, Twitter, Habrahabr and VKontakte) were used to investigate the distribution of Anglicisms. All selected online social networks have very wide topic coverage, user variety and ease of sampling a large dataset due to the public API. The source data, we used for training models and finding Anglicisms, is the following:

- VKontakte 11,426,003 Russian texts.
- Twitter 2,936,050 Russian texts.
- LiveJournal 10,000,000 Russian texts.
- Habrahabr 1,000,000 Russian texts.

To evaluate the proposed method we have used the following list of Anglicisms: we have combined the Dyakov dictionary with manually verified generated lists. The final dictionary contains 20,773 words. Subsequently, evaluation of our method will be performed based on this joint dictionary. The standard classification metric, F-measure, was used. It should be noted that due to the fact, that the algorithm cannot find Anglicisms that are missing in our corpus, we had to count only those words in the joint dictionary that have a frequency score of more than or equal to 30.

Table 1. Proposed method quality evaluated on different collections

Corpus	Method	True positive	False positive	Words of the joint dictionary in the corpus
VK+TW, LD ≤ 2	linear mapping	620	1,454	1,103
	SkipGram	323	235	
	linear mapping + SkipGram	823	1,638	
LJ, LD ≤ 2	linear mapping	506	323	4,321
	SkipGram	1,084	1,339	
	linear mapping + SkipGram	1,571	1,404	
Habr, LD ≤ 2	linear mapping	749	723	2,729
	SkipGram	534	139	
	linear mapping + SkipGram	1,060	554	

The corpus analysis on the material of blog-platform LiveJournal, that contains more than 20 million texts, has enabled us to detect Anglicisms. However, the intersection with the manually collected A. I. Dyakov dictionary of Anglicisms constitutes only 26%. At the same time more than 16,000 words were not presented in the LJ corpus at all. This allowed us to hypothesize—the Anglicism's usage was unevenly distributed among users of social networks. An alternative hypothesis was that users of LiveJournal did not use Anglicisms in their speech and writing at all. For hypothesis verification we have entailed statistical analysis of Anglicism distribution among users of VKontakte, Twitter, LiveJournal. additional analysis of user groups split by age and gender has been performed.

Distribution of Anglicisms among random users of VKontakte and LiveJournal social networks is shown in the picture 2 (a) and (b). Users who had at least 300 words in their texts were used to build the chart. The red color on the chart illustrates absolute number of Anglicisms in user texts. The blue color shows the number of unique Anglicisms. It shows that users from different social networks tend to apply Anglicisms in nearly the same proportion. Single Anglicisms were recognized between the 16th and the 25th word in user's speech irrespective of social network.

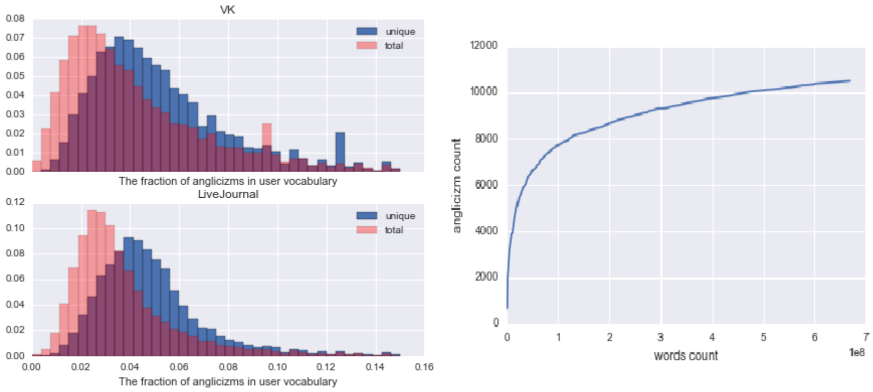

Fig. 2. The fraction of Anglicisms in a user's dictionary of **(a)** VKontakte **(b)** LiveJournal and the ratio of dictionary size to the size of the corpus with random search **(c)**

Frequency of Anglicism usage did not depend on gender, age or social network to be good evidence of fast adaptation of Anglicisms among users and their rapid integration into active dictionaries of online social network users researched. Analysis of VK user's age showed that users from age 12 to 35 tend to use new lexical items more actively than users from 40 to 70. The intersection between Anglicisms acquired from adults and teenagers was 0.62. Most frequent Anglicisms, used by one of the analyzed groups and almost never used by another is shown in table 2.

Table 2. Most frequent words, used only by teens and grown ups

Grown Up	Words frequency	Teen	Words frequency
маргарин (margarine)	1,602	шоурум (showroom)	957
бинго (bingo)	1,382	инст (inst)	311
компресс (compress)	1,009	мейк (make)	187
стак (stack)	969	трип (trip)	183
паприка (paprika)	851	спамить (spam)	158
форекс (forex)	729	хип-хоп (hip-hop)	149
майнинг (mining)	661	треш (trash)	147
компост (compost)	561	микроблейдинг (mikrobleyding)	147
тампон (tampon)	538	свитшот (svitshot)	140
тимьян (thyme)	510	кроссфит (crossfit)	139
рамблер (rambler)	492	пати (party)	137

Therefore, for actualization of dictionaries it's preferable to select young user's data, as Anglicisms used by people of the older generation are likely to be already contained in the dictionary. Furthermore, we can conclude that the hypothesis of uneven distribution of Anglicisms among Russian native speakers has been confirmed. Although observed social groups use different Anglicisms, the proportion of loan words in their speech is almost the same as it was in picture 2 (a) and (b). So we cannot say that teenagers use more Anglicisms, than grown-ups—observed words are different, but the proportion is the same.

Using the statistics of Anglicism usage, acquired from random user crawling, we have analyzed several dictionary formation strategies. Modelling Anglicisms by number of users that simultaneously use them, shows a strongly connected network with an average degree of 130. It assumes that we can significantly reduce a corpus size by focusing on users that have large amounts of Anglicisms in their speech. Our corpus formation algorithm had two steps, based on real crawling capabilities—(1) Get all texts (i.e. Anglicisms) of some user and (2) Get all users of some Anglicism. Step (1) supposes that we download all texts, include them to our corpus and proceed with the method described in section 2. To increase modelling speed, we made the assumption that Anglicisms can be found if they occur more than 30 times in the corpora. The number of Anglicisms can be estimated by multiplying the number of words by 0.35 to get the exact estimation (where 0.35 is the average F_1-measure quality of Anglicism detection by our method).

We took 100 users at each iteration; the user selection strategy was as follows:

- "Random"—select random users that were not included in to the previous iterations.
- "Rare A"—select users that actively use rare Anglicisms in their speech first.
- "Max A"—select users that use many Anglicisms in their speech first.
- "Max Lexic"—select users that have the richest vocabulary.

As we cannot see all user texts before we download them, we modeled our statistics on 100 randomly selected words written by this random user. This method simulates the situation when we observe comments associated with text already downloaded. We evaluated 1,000 experiments to get the mean statistic for each strategy. The resulting dictionary size ratio is shown in figure 3.

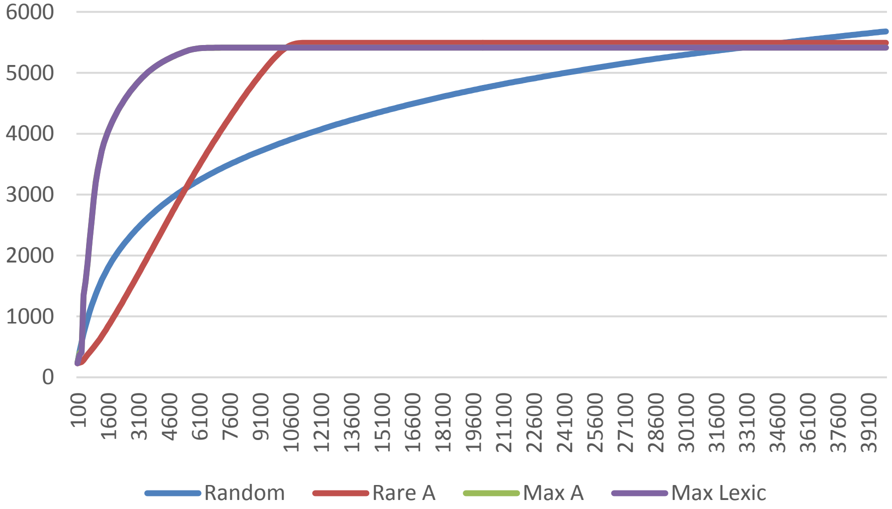

Fig. 3. The dictionary size to users downloaded ratio

As shown in figure 3, "Max A" and "Max Lexic" strategies give almost the same result. The dictionary size increases faster in the case of these strategies, although they are not able to get all Anglicisms found by random search strategy because some loan words stay separate from the rest of the vocabulary.

4. Conclusions

The following section breaks down three research contributions of this work and discusses their limitations. The linear mapping significantly increases total found borrowings recall and provides words missed by SkipGram model or naive translation. The resulting method is corpus dependent — it requires the same Russian word and its English analogue to be included into the corpus at least 30 times. The proposed method has satisfactory computational complexity that allows the researcher to verify hypotheses at Levenshtein Distance 2 or even more. Resulting recall at LD ≤ 2 is 0.74 that is significantly higher than all earlier observed results. The proposed method does not require precompiled dictionaries, however the use of the established dictionaries can be used to exclude old-fashioned Anglicisms and borrowings from other languages and giving researchers only contemporary, words unknown earlier.

Different social, gender, age groups, use different Anglicisms, although the percentage of loan words is nearly the same for all groups. Profile information should be used during the corpus formation as it increases the resulting Anglicism dictionary size. Teenagers use new lexical borrowings more actively than adult users, so the "New Anglicism Search" problem should be focused on a younger audience.

The best corpus formation strategy is the combination of random search and selection of rich vocabulary actors. First 5,000 users provide 95% of all Anglicisms contained in the corpus in this case.

Acknowledgement

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'.

References

1. *Balakina J.* (2011), Anglicisms in Russian and German Blogs, Frankfurt am Main: Peter Lang.
2. *Bylatcheva O. A., Safonkina O. S.* (2016), Internet Anglicisms: the development of English-Russian contacts today [Internet-anglicismy: rasvitie anglo-russkich yazykovych kontaktov na sovremennom etape] //Ogariov-Online, № 17 (82).
3. *Chachibaia N. G., Colenso M. R.* (2005), New Anglicisms in Russia, In and Out of.
4. *Chugunova E. I., Runtova N. V.* (2016), The origin and classification of Anglicisms in Russian language [Proishozhdenie i klassifikazia anglicismov v russkom yazyke], The modern tendencies of development of science and technology [Sovremennye tendencii rasvitiya nauki i technologij], № 10-6, pp. 135–138.

5. *Dyakov A. I.* (2016), Adaptational model of Anglicisms [Adaptazionnaya model anglicismov], Scientific researches: from theory to practice [nauchnye issledovania: ot teorii k praktike], № 3 (9), pp 245–255.
6. *Fenogenova A., Karpov I., Kazorin V. A.* (2016), General Method Applicable to the Search for Anglicisms in Russian Social Network Texts., AINL-FRUCT, IEEE p. 1–6.
7. *Gisle Andersen* (2005), Assessing algorithms for automatic extraction of Anglicisms in Norwegian texts, Corpus Linguistics.
8. *Janurik S.* (2010), The integration of English loanwords in Russian: An overview of recent borrowings, Studia Slavica, T. 55, № 1.,pp. 45–65.
9. *Kristiansen M.* (2013) Detecting specialised neologisms in researchers' blogs, Bergen Language and Linguistics Studies, T. 3, № 1.
10. *Leidig S., Schlippe T., Schultz T.* (2014), Automatic detection of Anglicisms for the pronunciation dictionary generation: a case study on our German IT corpus, SLTU, pp. 207–214.
11. *Mikolov T., Le Q. V., Sutskever I.* (2013) Exploiting similarities among languages for machine translation; preprint arXiv:1309.4168.
12. *Muraviev N. A., Panchenko A. I., and Obiedkov S. A.* (2014), Neologisms on Facebook, Dialog.
13. *Proshina Z.* (2008) English as a Lingua franca in Russia, Intercultural Communication Studies, T. 17., № 4., pp. 125–140.
14. *Serigos J.* (2016), Using distributional semantics in loanword research: A concept-based approach to quantifying semantic specificity of Anglicisms in Spanish, International Journal of Bilingualism, pp. 1367006916635836.
15. *Yaniv O.* (2016), Anglicisms in the Russian Language Based on-ing Borrowings.