On Difficulties of Compiling Parallel Corpus of Socio-Political Terms

Svetlana Manik*

Ivanovo State University, Ermaka street, 39, Ivanovo 153025, Russia

Abstract

The paper describes the process of compiling an online bilingual dictionary of socio-political terms and problems facing the lexicographer. It illustrates the challenges with the compilation of the parallel corpus of socio-political texts as the ideological evaluation influences the word choice in English- and Russian-speaking news media. The work began with the creation of LSP corpora of media texts on political and social issues in English and Russian and involved extraction of lexical information via the Sketch Engine tool. There is an attempt to describe a new format of a reference book - a rather large, semi-crowdsourced, mostly-data mined dictionary of English-Russian socio-political lexicon.

Keywords: Socio-political terms, parallel corpus, LSP, online bilingual dictionary, media text, ideological evaluation

1. Introduction

Words used by politicians, journalists, diplomats, political science students and, in general, non-professional people about political and social life are categorized in the field of ‘language and politics’ and described by the linguists as ‘language of politics’, ‘political/polticized language’ or ‘socio-political vocabulary/terms’.

* Corresponding author. Tel.: +79109850076; fax: +74932300291.
E-mail address: Svetlana_manik@yahoo.com
Political language as the generic term comprises “all types of public, institutional and private talks on political issues, all types of texts typical of politics as well as the use of lexical and stylistic linguistic instruments characterizing talks about political contexts” (Burkhardt, 1996). It includes talking about politics and political media language, as well as language of politics, i.e. terminology and rhetoric of politics with politicians acting as professionals.

Socio-political lexis is a term widely used in the Russian linguistics to denote a rather dynamic and multi-fielded vocabulary of political contexts. As there is a fast shift in the usage/preference of words and word combinations according to the significance of the socio-political event in the society ranging from annexation, terrorist act, Olympic Games, space exploration, sanctions war, Cold war to currency collapse, oil prices, social protests, cultural heritage, piracy, etc. henceforward socio-political terminology comprises terms of various spheres and it is relatively wider than ‘political language’.

Technological progress has made media political texts, the major source of information, numerous, diverse and available to a wider audience. Still the problems with understanding and interpreting of these texts and words are rather up-to-date. That is why there have appeared various online lists of political terms, political dictionaries, mono- and bilingual both printed and online editions. Previous researches have proved a great diversity of reference books registering words of politics and the words widely used in mass media to describe socio-political life on the Russian and foreign lexicographic markets (Manik, 2014). Several bilingual (Russian-English and English-Russian) dictionaries of politics, political and socio-political words have been published in Russia in the last two decades.

However, there still remains a need for multi-sided and updated description of the vocabulary used in English- and Russian speaking media covering events of the political life. The dictionary fills an important gap in bilingual LSP reference sources for Russian learners since there is still no contemporary bilingual (English-Russian) dictionary of socio-political vocabulary available (the latest Russian-English Dictionary by Tsarev (Царев, 2005) is dated 2005 and the latest edition of English-Russian Dictionary of Socio-Political Lexis is dated 1985 (Крупнов, 1985).

It may be correct to presume that the dictionary of LSP of politics and the one of LSP of socio-political life have different mission and target users: politicians, diplomats, journalists, students of Political Sciences, translators, scientists, etc. The focus of preparing dictionaries for a particular subject-field should be the needs of its user group in specific situation. This is catered for by the modern theory of dictionary functions and includes the introduction of subject-field components in dictionaries. Dictionary functions are either communication-oriented or cognition-oriented. In the case of the dictionary of socio-political vocabulary we mean a bilingual reference book to help the user get to grips with the political language through a detailed description of the entries.

The data have been logged in the carefully, manually and systematically populated database with some rough characterization of the entry meaning and its translation equivalent, its combinatorial behaviour in numerous examples (syntactical patterns, lexical collocations), indication of any stylistic, ideological, cultural, subject-field or other features. It started in 1997 and then it has been continued individually within the course Political English in Ivanovo State University.

The compilation began from LSP corpora of media texts on political and social issues in English and Russian and involved automatic extraction of lexical information from the corpus via the Sketch Engine tool. There is an attempt to describe a new format of a reference book, taking into account language data and NLP technologies already available, as well as the maturing technologies. It is conceptualized as an interactive web portal on English-Russian socio-political vocabulary where reliable information on all aspects of two particular languages is available. It is supposed to be a rather large, semi-crowdsourced, mostly-datamined dictionary of English-Russian socio-political lexicon.
The given paper defines peculiarities of the given terminology and describes the difficulties of compiling the parallel corpus of socio-political terms, a basic resource for contrastive and terminology studies, for research and development of translation systems, language analysis, teaching second language and translation didactics, semantic analysis and bilingual dictionary compilation. It is aimed at distinguishing technical and ideologically and cognitively grounded challenges during the lexicographic description of the LSP and ways to cope with them.

2. General Outline of Socio-Political Terminology

There are some features important for understanding the socio-political terminology (SPT):

- it is a culturally and ideologically meaningful lexicon; the notions terrorists, militants, separatists, rebels and ополченцы used in the same news context have different meanings and connotations in English- and Russian speaking societies;
- the ideology issue is important; it is understood beyond the political concept as a science/unity of ideas in a general sense and as a framework that is “assumed to specifically organize and monitor one form of socially shared mental representation, in other words, the organized evaluative beliefs — traditionally called ‘attitudes’ — shared by social groups” (van Dijk, 1996);
- evaluation aspect is essential since the language of politics, as W. Safire truly states in the prolegomenon of the updated and expanded edition of Safire’s Political Dictionary (2008), “is a lexicon of conflict and drama, of fulsome praise and fierce ridicule, of emotional pleading and intellectual persuasion. Color and bite permeate a language designed to rally support, to blast opposition, and to mold the minds of multitudes”;
- peculiarities of valued words are stipulated mostly by their functioning and lexical environment and not solely by the structure of the semantic meaning; socio-political lexis are very dynamic and much dependant on the lexical and grammatical environment;
- it denote numerous everyday and professional communication spheres and is wider than LSP of politics; some examples of the vocabulary: interrogation technique/method/tactics; on-going assault; breakaway republic; collapse; political clout; succumb to protectionism; banning ballistic missile testing; social unrest; to raise public-health alarms; constitutional amendment; to depose; deposed leader; ousted president; anti-government riot; ash cloud disruption; air travel chaos; interim government; acting president; Euromaidan; current blanket restrictions on airspace; bailout; stimulus plan; job cuts; low consumer confidence; restructuring of AIG’s operations; financial retrenchment; crusaders’ assault; the Arab Spring, to play a zero-sum game against smb.; megaphone diplomacy; to disperse anti-austerity protesters; austerity cuts; to be charged in the alleged manufacture of a homemade bomb; a national day of mourning; the flood-ravaged area; payout; to vow support; cyberbullying; suicide chat hotline; anti-cyberbullying push; allegabons; to provide immediate incentives for reformers; lack of transparency; Berkut; to hold vote-a-rama; fiscal cliff; fracking; Obamaloney; super pac; legitimate rape; the apology tour; job creator; the Right Sector; collateral damage; the war on terror; Occupy Wall Street; etc.;
- it is widely and commonly used in mass media.

Thus socio-political language is specific due to its composite nature: while it primarily relates to the field of politics and political science, it also stretches out to very different fields of human activity and changes rather dynamically.

The connotational and ideological difference between the word/word combination and its translation equivalent is very important. The terminology of socio-political life denotes conflicts and contradictions in the society, multifacetedness and critical character of numerous social processes and reforms, ethnic discords, wars and ‘cleansing’, political ‘battles’ for the votes and power, religions and gender relations, social and economic debates on unemployment, financial crisis, mortgage, credit crunch, etc.. Thus it has its peculiar (ideological, social, gender, cultural, etc.) features and characteristics. It is stipulated by the identity of every community and consequently it is reflected in the language. Simultaneously the lexis reproduces the government stand of the given country.
bilingual dictionaries (English-Russian) do not provide such comments on the cultures’ differences, though it might be rather useful.

3. Compilation of Parallel Corpus of Socio-Political Terms

3.1. Resources of Parallel Corpus

A parallel corpus is defined as a corpus comprising source texts and their translations. There is a great Sketch Engine Parallel Corpus tool available with 1,441,844,046 tokens in OPUS2 English and 381,468,257 tokens in OPUS2 Russian (Fig.1).

By now we work with two separate corpuses that later ideally to be transferred into a parallel corpus. The Corpus of Socio-Political Terms (CSPT) in English currently contains president’s addresses (Putin and Obama), official statements, reports before the Senate/Council of Federation, State of the Union Address, inauguration speeches, texts of G20 declarations, international protocols, agreements, interviews of the politicians, lawmakers, public leaders, texts of official documents, media publications with issues published both in Russian and English (BBC, Russia Today, Voice of America), etc. Another corpus is named Translation Corpus of Socio-Political Terms in Russian and it consists of similar texts as the first one but translated into Russian. The resources are mainly written texts published in the English- and Russian-speaking media in the period of 2012 till the present time. By now the corpus comprises over 300 texts and translations.

The source political context texts are rather diverse, that is why to discuss parallel text alignment and understand alignment problems it is necessary to point out some translation characteristics according to the dependency between the original text of the political context and its translation:

- **translated text completely substitutes the original text in the target language.** This is the case of literary translations when the readers usually choose to read only one version of them. These are the texts of the official documents, agreements, protocols, etc.

For example, the texts of G20 Leaders Declaration are usually published translated into Russian on the Kremlin website - http://kremlin.ru/ref_notes/1245. It looks as follows:
6. Euro Area members of the G20 will take all necessary policy measures to safeguard the integrity and stability of the area, improve the functioning of financial markets and break the feedback loop between sovereigns and banks. We look forward to the Euro Area working in partnership with the next Greek government to ensure they remain on the path to reform and sustainability within the Euro Area.

6. Участники еврозоны из числа государств «Группы двадцати» примут все необходимые меры политики, чтобы защитить целостность и стабильность региона, улучшить функционирование финансовых рынков и вырваться из замкнутого круга взаимозависимости между суверенными структурами и банками. Мы надеемся на плодотворное партнерство между еврозоной и будущим правительством Греции для обеспечения неизменности курса на реформы и устойчивость еврозоны.

Other resources are presented on the US State Department site, official website of Russia’s G20 presidency, on official site of major international organizations and NGOs and on some news agencies sites.

The source text and its translation are rather identical in the structure, style and terminology.

- translations coexist in time and space. This is the case of bilingual literary editions (where the reader will probably compare the texts on both languages). There are publications of news agencies on the same issue in two languages. The topic is the same, the fact mentions are relevant, but the structure and presentations may differ.

For example, in January 12, 2015 RT published two articles: Central Command Twitter account apparently hacked by CyberCaliphate and Сторонники «Исламского государства» взломали аккаунт центрального командования США в Twitter. As it is evident from the headings the theme is identical, but the English notion ‘CyberCaliphate’ is interpreted in Russian as ‘сторонники «Исламского государства»’ (advocates/supporters of the Islamic State), it gains another colouring and connotation. Thus media texts and their translations cause some difficulties while compiling parallel corpus due to alignment problems.

- translations are used for the same purpose as the original, and work in a symmetrical way. This is the case for institutional documents of the European Union, United Nations Organisation and other multilingual institutions. The purpose of these texts is informative, rather than manipulative, if compared with the previous group.

- translations are made partially and not necessarily targeted at the same objective. For example, Obama’s speeches are translated into Russian by parts, almost 2/3 are missed (http://iipdigital.usembassy.gov/st/russian/texttrans/2015/01/20150122313106.html#axzz3VOGWs814). This is the case when the target audience is different and the message is revealed through different means and techniques. In this case the CSPT includes only parts of the source text that have the translation.

It is very important to note the objective of the translation of the political context texts since the political communication is targeted at reaching some goal and executing some strategy (informative, manipulative, etc.). It may vary from maintaining the same communication purpose as the original, keeping the original structure and form of the language to transmitting essentially the same/different message.

To sum up, the diversity of source texts and their translations bring rather crucial problems as parallel text alignment problems are highly dependent on the stated above facts. On the one hand, in some cases translations cannot be viewed as parallel corpora since the translator often changes the order of sentences and add or omit some content and comments as soon as they maintain the basic idea behind the text. On the other hand, translations may be the best type of parallel corpora for alignment. As this type of parallel corpora is normally composed of institutional documents with laws and other important information, translation is done accurately, so that no ambiguities are inserted in the text, and they maintain symmetrical coherence.
3.2. Technical Difficulties of Parallel Corpus Compilation

The last decades have been rather productive and eventful because of the information technology development and technological progress in lexicography in particular. Various tools such as Word Sketch (Kilgarriff and Tugwell, 2002) and TickBox Lexicography (Kilgarriff et al., 2010) have been designed as part of corpus query systems to help lexicographers work with the language and build larger and larger corpora, but their design and purpose still requires lexicographers to select and transfer relevant corpus information to the dictionary writing system, like iLEX (Erlandsen, 2004). Thus the lexicographer’s role has changed to choosing and evaluating the choices suggested by the computer out of the corpora. Such a new approach to lexicographic work, in which the lexicographer is seen more as a validator of choices made by a computer, was envisaged by Rundell and Kilgarriff (2011). In terms of socio-political corpus the lexicographer’s role is more complicated.

Still a parallel corpus is difficult to create as it requires a non-straightforward process of alignment, i.e. pairing of source and target ‘equivalent’ segments, on a sequential basis, thus special software. While working out a concept of the parallel corpus there has been a dispute whether the given corpus is rather a comparable one, a corpus comprising not-identical, but similar in their content texts in one or more languages. Traditionally comparable corpus is defined as a set of texts, collected independently from their languages and connected on the basis of their content, domain and communicative function. On the one hand, the content seems to be the same: particular events of the socio-political life, speeches, agreements, etc. On the other hand, there is a great difference in the presentation and interpretation of these events depending on the official stand or point of view of a journalist, editor-house, news agency, etc., and media coverage much depends on it. That is why the following word-combinations used in the context of Crimea annexation/joining – присоединение Крыма are widely spread in mass media:

- Russia’s military led annexation of the Crimean peninsula from Ukraine (CNN)
- Citizens in the disputed peninsula of Crimea voted to leave Ukraine in a hastily arranged
  Putin’s takeover of Crimea (CNN)
- Ukraine’s breakaway peninsula (CNN)
- But dealing with such a blatant land grab on its eastern flank won’t be anywhere near as quick and easy
  for Europe’s 28-member union (CNN)
- annexation of Crimea (the Economist)
- pro-Russian referendum in Crimea had no legitimacy (Economist)
- Crimea’s choice of joining Russia (RT)
- the right to self-determination for the peninsula’s population

They definitely reflect the ideological opinion of different sides and are of great value for the lexicographer to register all word-combinations as collocations and contexts in the dictionary of socio-political lexis and present the current state of the language development. But they are unlikely to come across with in the parallel corpus as the official documents, agreements, protocols etc. stipulate a neutral term ‘annexation’ – ‘аннексия’. It is also interesting to note that the understanding of this diplomatic term differs in English and Russian science (Manik, 2013).

Meanwhile Sketch Engine Parallel Corpus provides a concordance in the Parallel Corpus to the query ‘присоединение (joining) – annexation’ (Fig.2).
But the matching examples illustrate other historical events dated back to 1823, 1981, 1989, 2000, etc. of Golan Heights, Alaska, Israel, Cuba, Cyprus, Estonia as well as in the query ‘annexation – присоединение’. Thus the data do not cover the current situation, do not reflect the modern contexts and do not differentiate the semantic difference, connotations, between the translation equivalents in Russian присоединение – присоединение.

Technically we hope to make use of Sketch Engine Parallel Corpora tool in the future and build the up-to-date Socio-Political Parallel Corpus aligned both at the paragraph and sentence level. Two present corpora and a running project on the bilingual dictionary of socio-political terms will ensure a higher accuracy of the alignment.

3.3. Strive to Be Neutral while Compiling and Processing the Corpus Findings

As it has been proved by the research (Manik, 2013) the ideological evaluation plays a significant role in the socio-political vocabulary as, on the one hand, it influences the word choice by the speaker, even within one language (for example, separatist – rebel – terrorist – combatant; annexation – joining – land grab); on the other hand, evaluation becomes explicit in/through the context. It sums up the idea of the necessity to study the contexts of the political language to understand the real and true meaning of the words.

It is the fact that many words and word combinations from the author’s corpora are not in the modern political dictionaries as media texts provide rather diverse and stylistically coloured vocabulary (mostly with the help of metaphors, metonymy and allusions). Besides it is challenging to update the dictionary or the corpus as fast as the socio-political developments occur and replace each other in mass media. For example, Sketch Engine English corpus enTenTen12 is dated 2012. In terms of significance of the examples, time-tested, they do not meet Atkins and Rundell requirements: “naturalness and typicality, informativeness and understandability” (Atkins, Rundell, 2008:458) because these contexts do not demonstrate the current preferences of the speakers.

It is of vital importance to include various authoritative and reliable resources in the corpus to cover the diversity of points of views, ideological stands of different communities and not to prescribe the word choice. Consequently there are examples from official government resources, from documents, diplomatic protocols, from pro-Russian, pro-American, pro-European, etc. mass media.

Thus the lexicographer’s role, to our mind, has changed to choosing and evaluating the choices suggested by the computer out of the corpora in order to provide full coverage of the language development and keep the neutral
ground during further lexicographic description of the words.

4. Applications of the Bilingual Corpus and Future Development

Compilation of the Parallel Corpus of Socio-Political Lexis has from the very beginning been seen as a project that would help to create a bilingual LSP dictionary, a rather large, semi-crowd-sourced, mostly-datamined dictionary of English-Russian socio-political lexicon, and make regular information update possible. It is aimed at mostly Russian-speaking users with the intermediate or advanced level of English who are interested in Politics, socio-political life of the country and the world: students, teachers, politicians, diplomats, translators, etc.

The following information is registered in the suggested dictionary microstructure:

- translation equivalent(s) with additional comments in brackets concretizing the key point of the notion;
- cultural reference (Interesting! Section) including the detailed description of the entry and some brief current outlook on ideological perceptions or interpretations in the media;
- illustrative examples with the source indication to show the entry in the context of usage. The quotation is selected from the corpus to confirm and add the semantics of the translation equivalent and the culture data. This section should be updated regularly and it contains pretty many examples from different resources;
- visuals (if necessary) borrowed from flickr.com to illustrate the notion;
- collocations (relatively close in the meaning to the entry) and different word combinations used in the same context. They will be selected form the parallel corpus and will widen the user’s vocabulary and help to understand the difference in the variety of media interpretations;
- each word or word combination has a cross-reference and can be read as a separate dictionary article.

On the one hand, such a detailed description of the entry may seem exhaustive. On the other hand, user’s needs survey has proved the demand on additional data presented by sections/paragraphs with the help of lexicographic semiotics. It’s is worth noting that the electronic format of the dictionary allows a user to choose the options or limit the reference to the translation equivalent.

We believe that the log file concept should also be implemented as a potential source of information on user’s needs, on the one hand, and as improving the data presentation in the portal, on the other hand.

The use of a better search system which includes the possibility of a direct search for synonyms, collocations, antonyms, word formations, grammar items etc. will provide a much more precise way of obtaining knowledge about real dictionary use. On the basis of such data, it will be possible to prepare much better internet dictionary of socio-political vocabulary.

5. Conclusion

Lexicography nowadays is facing great changes in the data analysis and processing domains. Automation can be implemented in many aspects of lexicographer’s work saving time and money and enlarging the amount of information systematized. Nonetheless, some tasks, especially anything connected with semantics, are still the challenges the linguists have to deal with.

In terms of Parallel Corpus of Socio-Political Terms as discussed in the paper there exist several difficulties stipulated technically and semantically. On the one hand, its compilation is still in process as with the help of Sketch Engine we are working on two separate corpora in English and Russian which later should make one parallel corpus aligned both on sentence and paragraph levels. On the other hand, ideological peculiarities and diversity of the notions and their translations in the political context texts complicate and harden the process.
References


