Documentation of Data Collection

The President's Words:

A Term Frequency Analysis of Putin's and Medvedev's Statements in Official Kremlin Transcripts

Daniel Marcus

dekoder, www.dekoder.org

dm@dekoder.org

Content

This dataset consists of two parts:

(1) A database of more than 10,000 transcripts published on the the official website of the President of Russia since 1999, December 31st (last update 2021, March 1st).
(2) A frequency analysis of the terms used by president Putin and president Medvedev in these documents between 2000 and 2020.

This is the data used in the word counting tool at https://putin.dekoder.org/.

Keywords: Russia; Kremlin; Putin; Medvedev

Context of the data

Vladimir Putin was elected to the Russian presidency for the first time in March of 2000. He has been in power now for twenty years. Twenty is a lot of years. And a lot has happened: terrorist attacks, economic crises, the Russo-Georgian War, the annexation of Crimea and the war in eastern Ukraine, sanctions on Russia and Russia's countersanctions. There have been extensive reforms, the inaugural sessions of five newly elected parliaments, and large-scale demonstrations.

The dekoder special "20 Years of Putin" (https://putin.dekoder.org/words) is an attempt to decode Putin. For this purpose, we have developed a tool that analyses texts from the official website of the President of Russia to generate a graphic representation of the frequency of word use by Putin (2000–2008 and 2012–2020) and Dmitry Medvedev (2008–2012). We analysed more than 10,000 Kremlin publications to determine how often which terms appear

in them. In preparing and depicting the data, we based our procedure largely on that used by our colleagues at Zeit Online for their project 70 Jahre Bundestag – Darüber spricht der Bundestag.

What data did we use?

The analysis is based on raw data pulled from more than 10,000 transcripts published on the official website of the Kremlin between 1 January 2000 and 31 December 2020 (Russian: http://kremlin.ru/events/president/transcripts, English: http://en.kremlin.ru/events/president/transcripts). These include official addresses given by the Russian president and transcripts of meetings and interviews, as well as other kinds of texts, such as op-eds by the president that appeared in various newspapers.

Dmitry Medvedev served as Russia's president from 7 May 2008 to 6 May 2012, with Vladimir Putin assuming the post of prime minister. Thus, the data for this period which is known as "castling" relate to statements made by Medvedev. From 7 May 2012 onwards, they once again relate to Putin's words. We have labelled the years in the chart accordingly.

What procedure did we follow?

The transcripts posted on kremlin.ru also contain the speech of people other than the Russian president – people who attended meetings with the president, interviewed him, etc. Although we did our best to filter this speech out, we cannot absolutely rule out the possibility of slight distortions resulting from content of this kind, because the transcripts do not always indicate a change of speaker in a uniform manner.

The first thing we did to prepare the data for analysis was to chop the filtered transcripts up into individual words, known as tokens. Then we removed all of the "stop words" from the token list – i.e. words like "and" (и), "so" (так) or "only" (только), which have no particular relevance for the analysis.

Individual terms (especially in Russian) can occur in a variety of forms (газета, газеты, газете, газету, …), so the next step was to standardise all the variants, i.e. change them all to their dictionary form, or lemma. In computational linguistics, this step is called lemmatisation. We used an algorithm developed by the Russian search engine provider Yandex for this. (For the English version: StanfordNLP LemmaProcessor.)

We also searched the data for words that occur in two or three-word strings (known as n-grams) with particular frequency, because we were interested in combinations of words like "artificial intelligence" (искусственный интеллект) or "Great Patriotic War" (Великая Отечественная Война), as well as in individual terms.

The last step was to count the number of times that the words and word combinations appear in the data associated with each individual year. To ensure that differences in the volume of material published in different years would not distort the results, we set up the tool to chart relative rather than absolute frequency; i.e. it shows the frequency with which a word or a combination of words appears per 100,000 words in a year.

What else should users keep in mind?

Like the original documents, the data may contain misspelled words. To keep the dataset to a manageable size, only terms occurring at least three times over the entire period are shown.

The data for the German and Russian versions of this tool were derived from the Kremlin's Russian-language publications; for the English version of the tool, we used the English-language publications. There may be differences between the English and Russian versions of a chart, since the English-language Kremlin site posts somewhat fewer documents and because the translations can sometimes differ from their source texts, and can also be prepared using different spelling standards ("modernisation" vs. "modernization").

Data

(1) The kremlin.ru transcripts are saved as Elasticsearch index dumps in zipped JSON-files, one for each language (kremlin_transcripts_ru.json and kremlin_transcripts_en.json). Each transcript document includes the following extracted information:

- date (of the original publication)
- kremlin_id (the internal document ID from the kremlin.ru website)
- persons (a list of tagged persons)
- place
- tags
- teaser
- title
- transcript (technically a concatenated string of all text paragraph's contents)

(2) The results of the term frequency analysis are saved in a zipped CSV table (one for each language). It includes the following information:

- term
- year
- count (how often does the term occur in that particular year)
- freq (how often does the term occur in that particular year per 100,000 words)
- total_count (how often does the term occur in all years [2000-2020])
- total_freq (how often does the term occur in all years [2000-2020] per 100,000 words)
- ngram (is this a single word or does it consist of 2 or 3 words?)
- tf_idf_score (term frequeny – inverse document frequency: reflects how "special" the term frequency for that particular year is compared to the other years)

Using and analysing the data

dekoder developed a free tool that analyses texts from the official website of the President of Russia to generate a graphic representation of the frequency of word use, available at https://putin.dekoder.org/words.