transcript.duma.gov.ru_ru_2024

Corpus based on the Russia's Duma website (in Russian, 2006-2023)

Giorgio Comai (OBCT/CCI)

2024-09-16

Scope of this corpus

This corpus is based on all transcripts of Duma sessions as published on the official website transcript.duma.gov.ru as it was available online in early 2024. All text of session transcripts and voting is extracted as such, e.g. without differentiating by speaker, or parsing vote results.

Note

Users of this dataset should be aware that Russia's Duma makes available these contents (and more) through a dedicated API available at the following address: http://api.duma.gov.ru/. This however requires to obtain an API after requesting it. The data available through API for the period 1994-2021 have previously been extracted and are available on *Discuss Data*:

> dekoder.org (2021): Duma Speeches: A Term Frequency Analysis – Russian State Duma Transcripts 1994–2021, v. 1.0, Discuss Data, https://doi.org/10.48320/FB52DAC2-66E3-47A3-86C5-B2A3DADF41BF

Summary statistics

Dataset name: transcript.duma.gov.ru_ru_2024

Dataset description: all transcripts published on transcript.duma.gov.ru

Start date: 1994-01-11

End date: 2023-12-15

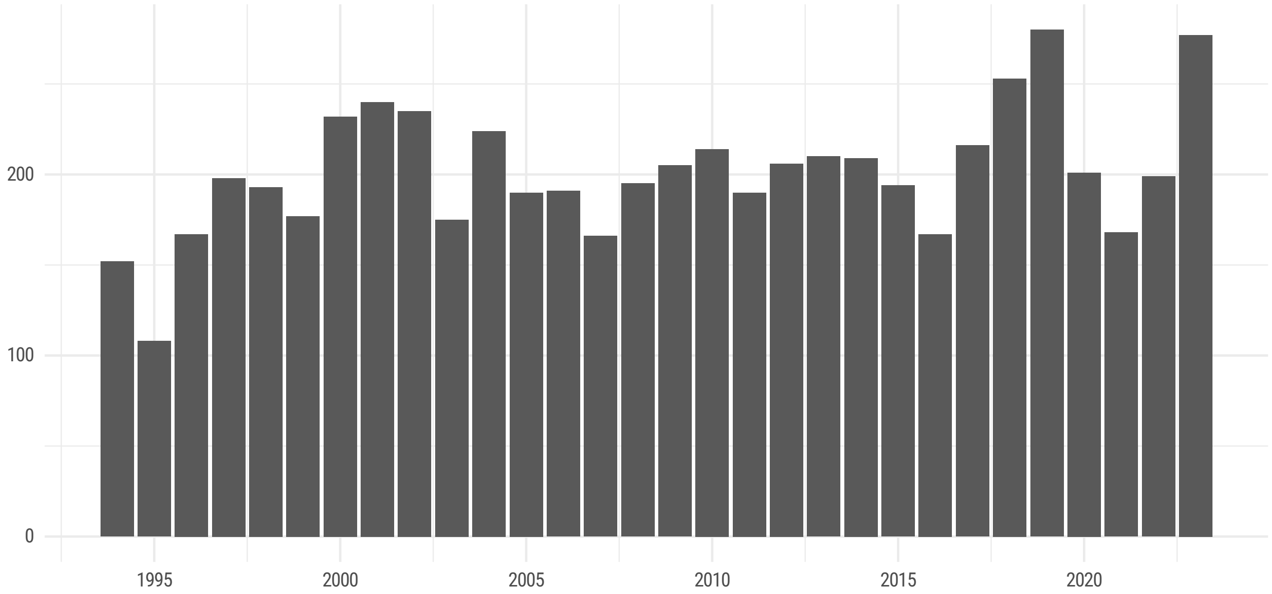
Total items: 6 032

Available columns: doc_id; text; title; date; url_id; url

License: see details

Number of items per year published on transcript.duma.gov.ru

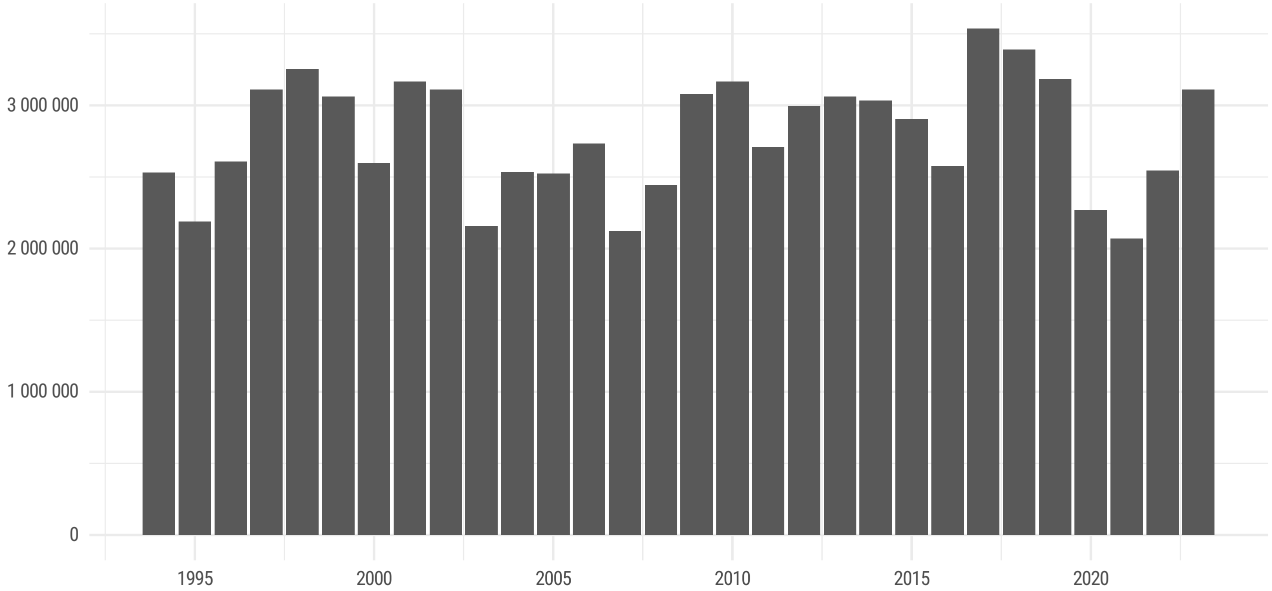
Based on 6 032 items published between 11 January 1994 and 15 December 2023



Source: Giorgio Comai / tadadit.xyz / transcript.duma.gov.ru_ru_2024

Number of words per year published on transcript.duma.gov.ru

Based on 6 032 items published between 11 January 1994 and 15 December 2023



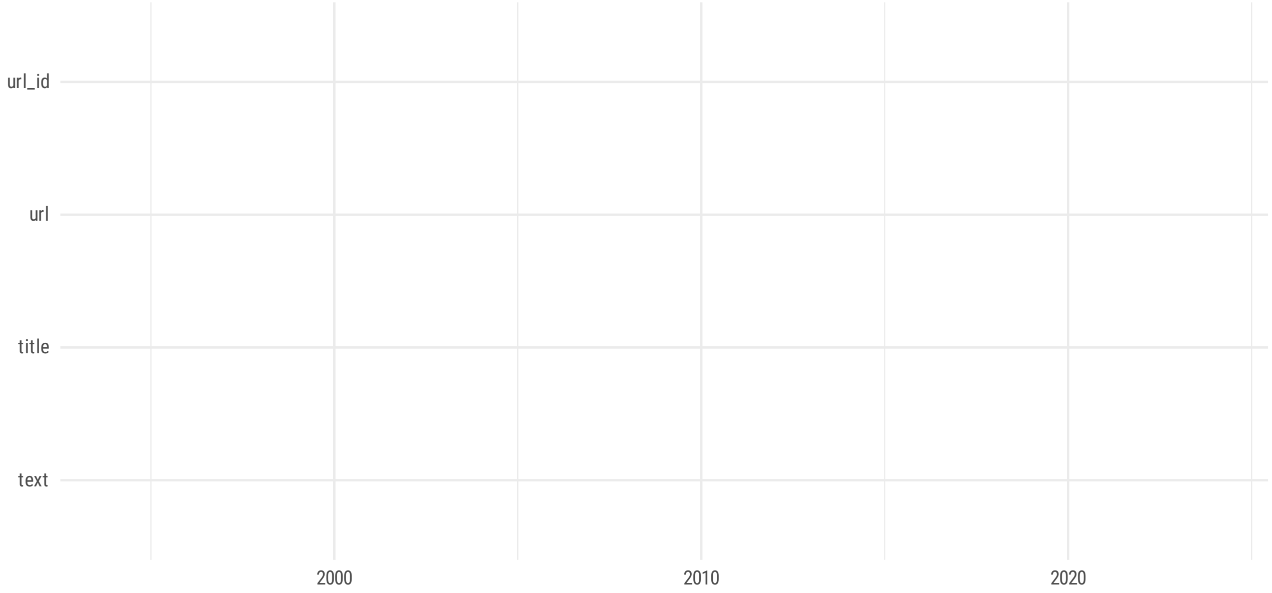
Source: Giorgio Comai / tadadit.xyz / transcript.duma.gov.ru_ru_2024

Missing data

field	present	missing	missing_share
doc_id	6 032	0	0.0%
text	6 032	0	0.0%
title	6 032	0	0.0%
date	6 032	0	0.0%
url_id	6 032	0	0.0%
url	6 032	0	0.0%

Missing metadata on transcript.duma.gov.ru

A thin line represents an empty field



Source: Giorgio Comai / tadadit.xyz / transcript.duma.gov.ru_ru_2024

Narrative explanation of how this corpus has been created

Rather than by querying the archive, this dataset has been created by all urls based on the observation that the page of each url is made of a numeric identifier, e.g.:

`http://transcript.duma.gov.ru/node/1234/`.

Urls that returned missing pages were discarded.

License information

The section of Duma's website dedicated to transcripts does not have a dedicated page with terms of use or licensing information. Its footer includes a generic copyright notice, claiming copyright.

> © Государственная Дума Федерального Собрания Российской Федерации, 2024

The about page of the main Duma website, of which this transcripts section is ultimately part, includes a page "On the use of information" ("Об использовании информации"), which clarifies the permissive conditions for re-publishing contents used on the website. Even if it does not include reference to specific license, it unambiguously states that contents can be published anywhere, without any sort of limitation, with the only condition being that a link the original source must be included. It appears that the same terms of use should apply also to the "transcripts" section of the website.

> Все материалы официального сайта Государственной Думы Федерального Собрания Российской Федерации могут быть воспроизведены в любых средствах массовой информации, на серверах сети Интернет или на любых иных носителях без каких�либо ограничений по объему и срокам публикации. Это разрешение в равной степени распространяется на газеты, журналы, радиостанции, телеканалы, сайты и страницы сети Интернет. Единственным условием перепечатки и ретрансляции является ссылка на первоисточник. Никакого предварительного согласия на перепечатку со стороны Аппарата Государственной Думы не требуется.

The contents of this dataset - "transcript.duma.gov.ru_ru" - are distributed within the remits of this license. To the extent that it is possible, the dataset itself is also distributed by its creator, Giorgio Comai under the Open Data Commons Attribution license (ODC-BY).

Funding and disclaimers

This dataset and accompanying materials have been produced within the scope of a project conducted with the support of the Unit for Analysis, Policy Planning, Statistics and Historical Documentation - Directorate General for Public and Cultural Diplomacy of the Italian Ministry of Foreign Affairs and International Cooperation, in accordance with Article 23 – bis of the Decree of the President of the Italian Republic 18/1967.

The views expressed are solely those of the authors and do not necessarily reflect the views of the Ministry of Foreign Affairs and International Cooperation.