

# Text-mining the post-Soviet web

**Giorgio Comai**  
*Dublin City University*  
*@giocomai*

# Structure of the presentation

- Creating textual datasets from the web
- Why bother?
- Some examples of results
- If it was easy enough, would area studies researchers use this approach?

# Creating datasets

The screenshot shows the homepage of the Ministry of Foreign Affairs of Abkhazia. At the top, there's a navigation bar with links to Home, Travel to Abkhazia, Ministry, Foreign Policy, Entrance to Abkhazia, and Press. Below the navigation is the official seal of Abkhazia. The main content area features several news items with small thumbnail images:

- 13.03.2015: Apsny Foreign Minister V.A. Chirikba sent notes of condolences to the Minister of Foreign Affairs of the Republic of Armenia and Turkey.
- 14.03.2015: The Foreign Ministry of the Russian Federation hosted a meeting of the participants of the Geneva discussions on security and stability in Transcaucasia, which was attended by representatives of the Russian Federation, the Republics of Abkhazia and South Ossetia.
- 11.03.2015: Meeting of the Minister of Foreign Affairs of the Republic of Abkhazia V.A. Chirikba and the Minister of Foreign Affairs of Russia Sergey Lavrov.
- 10.03.2015: During his visit, the working visit of the Minister of Foreign Affairs of the Republic of Abkhazia V.A. Chirikba and the Minister of Foreign Affairs of the Russian Federation Sergey Lavrov. A thorough exchange of views on issues of developing bilateral, regional and international agenda took place.
- 11.03.2015: Working visit by the Minister of Foreign Affairs of Abkhazia.
- 09.03.2015: The working visit of the Minister of Foreign Affairs of the Republic of Abkhazia V.A. Chirikba and the Minister of Foreign Affairs of the Russian Federation will be held from 10 to 12 March.
- 08.03.2015: During his visit on March 11, bilateral talks with the Minister of Foreign Affairs of the Russian Federation Sergey Lavrov will be held.
- 07.03.2015: On the meeting with M.A. Balashov.
- 06.03.2015: On March 3 in the Ministry of Foreign Affairs of Abkhazia was held a meeting of the Minister of Foreign Affairs of the Republic of Abkhazia V.A. Chirikba and the Counselor-Envoy of the Embassy of the Russian Federation in the Republic of Abkhazia G.P. Klykov.
- 27.02.2015: On February 26, the Head of the International Committee of the Red Cross (ICRC) in Abkhazia, Mr. Georges Mandarachvili, met with the Minister of Foreign Affairs Vachchel Chirikba with the Head of the ICRC mission in Abkhazia.
- 25.02.2015: On February 26, in the MIA of Abkhazia was held a meeting of the Head of the International Committee of the Red Cross in Abkhazia with the Head of the ICRC mission in Abkhazia Georges Mandarachvili.
- 23.02.2015: Comment of the Ministry of Foreign Affairs of Abkhazia.
- 20.02.2015: From 10 to 15 February 2015 was held a working visit of the Abkhazian delegation composed of the Head of the International Committee of the Red Cross in Abkhazia, the representative of the Initiative group of the Republic of Georgia, the Republic of Abkhazia in Turkey and Perpetual Representative of the Republic of Abkhazia in Russia. The Head of the International Committee of the Red Cross in Abkhazia, the Head of the International Committee of the Red Cross in Turkey and the Perpetual Representative of the Republic of Abkhazia in Russia during the visit visited meetings of humanitarian, economic and political nature. This caused a flurry of indignation from the Georgian officials.
- 19.02.2015: Comment of the Abkhaz Foreign Ministry.
- At the 7th International Tourism Fair "Salon Tur" in Belgrade the state symbol - the flag of the Republic of Abkhazia was displayed. The official member - the Chamber of Commerce of Abkhazia, The Ministry of Foreign Affairs of the Republic of Abkhazia and the Ministry of Culture of the Republic of Abkhazia attempt by the Georgian side and its European allies" to level the value of the Abkhazian flag and the flag of the Republic of Abkhazia and give the participation of the Abkhaz delegation in the exhibition purely commercial nature. Georgia's ongoing attempts to exert pressure on the Abkhaz representatives abroad deserve strong condemnation.

Below the news items is a contact form for "Consular Service" with fields for Name, Surname, Email, and Message. To the right is a sidebar with links to "About Abkhazia", "Ministry", "Foreign Policy", and "Consular Information".

This screenshot shows a specific news item from the website. The news item is dated 28.05.2014 and has a red circle drawn around it. The text of the news item is:

A telephone conversation took place between Viacheslav Chirikba and Grigory Karasin

During the telephone conversation of the Abkhazian Foreign Minister Viacheslav Chirikba with the Deputy Foreign Minister Grigory Karasin, V. Chirikba acquainted G. Karasin with the prevailing situation in Abkhazia.

Karasin "expressed concern about the worsening political situation in Abkhazia, where on May 27 actions were held organized by the opposition." "The Russian side with attention and concern follows the events in the friendly Republic and considers it important that the political processes there evolved exclusively in the legal course," - said the Deputy Minister of Foreign Affairs of Russia.

Below the news item is a link "Возврат к списку" (Return to list). At the bottom of the page are social media sharing buttons for Facebook, Twitter, Google+, and Vkontakte, along with a LiveJournal button. The footer contains links to "Government & Institutions", "Embassies", "Mass media", and "Extra".

Page footer: Работа сайта: Mukhus | MINISTRY OF FOREIGN AFFAIRS REPUBLIC OF ABKHAZIA | E-mail: info@mfaapsny.org; Tel: + (840) 226-70-69; Adress.: Sakharova, 33

```
> meta(corpus[[958]])
```

Metadata:

```
author      : Abkhazia's MFA
datetimestamp: 2014-05-28
description  : character(0)
heading      : A telephone conversation took place between Viacheslav Chirikba and Grigory Karasin
id          : ABK2461
language     : en
origin       : http://mfaapsny.org/en/information/?ID=2461
```

```
> corpus[[958]]
```

<<PlainTextDocument (metadata: 7)>>

28.05.2014

A telephone conversation took place between Viacheslav Chirikba and Grigory Karasin  
During the telephone conversation of the Abkhazian Foreign Minister Viacheslav Chirikba with the Deputy Foreign Minister Grigory Karasin, V. Chirikba acquainted G. Karasin with the prevailing situation in Abkhazia.

Karasin "expressed concern about the worsening political situation in Abkhazia, where on May 27 actions were held organized by the opposition." "The Russian side with attention and concern follows the events in the friendly Republic and considers it important that the political processes there evolved exclusively in the legal course," - said the Deputy Minister of Foreign Affairs of Russia.

# Export, archive, etc.

B	C	D	E	F	G	H	I
nameOfP	nameOfW	dates	articles	titles	language	articlesLinks	articlesTxt
MFAs	Abkhazia	2012-04-16	1	On April 16-18- the Minister of Foreign Affairs of the Republic of Abkhazia will take part in the Geneva discussions	en	<a href="http://mfaapsny.org/en/information/?ID=44">http://mfaapsny.org/en/information/?ID=44</a>	16.04.2012 On April 16-18, the Minister of Foreign Affairs of the Republic of Abkhazia will take part in the Geneva discussions. On April 16-18, the Minister of Foreign Affairs of the Republic of Abkhazia will take part in the Geneva discussions. During his stay in Moscow, Viacheslav Chirkba will take part in the Geneva discussions. During the visit, the Foreign Minister of Abkhazia also plans to visit
MFAs	Abkhazia	2012-04-17	2	The Minister of Foreign Affairs of the Republic of Abkhazia, Viacheslav Chirkba met with the UNDP Resident Representative Jamie McGoldrick	en	<a href="http://mfaapsny.org/en/information/?ID=47">http://mfaapsny.org/en/information/?ID=47</a>	17.04.2012 The Minister of Foreign Affairs of the Republic of Abkhazia, Viacheslav Chirkba met with the UNDP Resident Representative Jamie McGoldrick. On April 17, 2012 in Moscow, the Minister of Foreign Affairs of the Republic of Abkhazia, Viacheslav Chirkba met with the UNDP Resident Representative Jamie McGoldrick. At the meeting a wide range of issues of mutual interest were discussed.
MFAs	Abkhazia	2012-04-18	3	On April 18- 2012 a press conference of the Ministry of Foreign Affairs of the Republic of Abkhazia was held	en	<a href="http://mfaapsny.org/en/information/?ID=53">http://mfaapsny.org/en/information/?ID=53</a>	18.04.2012 On April 18, 2012 a press conference of the Minister of Foreign Affairs of the Republic of Abkhazia was held. On April 18, 2012 a press conference of the Minister of Foreign Affairs of the Republic of Abkhazia was held. Journalists were interested in a number of issues, in particular, the Geneva discussions. The head of the Abkhazian Foreign Ministry, told the journalists about the Geneva discussions. According to him, the Geneva discussions are the only place where the parties can discuss the issue of Georgia. Journalists at the press conference also touched the issue of Georgia. The visit of the delegation of the Ministry of Foreign Affairs of Abkhazia to Geneva was also mentioned. The press conference lasted more than an hour.

- Spreadsheet
- Folder of txt files
- All articles within a single txt file
- Document-term matrix
- ...

 2012-09-20 - AbkhaziaMfa - 88 - The Ministry of Foreign Affairs of the Republic of Abkhazia sends a congratulatory letter to the President of the Republic of Georgia.txt
 2012-09-21 - AbkhaziaMfa - 89 - Information by the Foreign Ministry of the Republic of Abkhazia about the Process of Geneva discussions.txt
 2012-09-24 - AbkhaziaMfa - 90 - Viacheslav Chirkba met with the UNDP Resident Representative Jamie McGoldrick.txt
 2012-09-25 - AbkhaziaMfa - 91 - Deputy Foreign Minister Irakli Khintba interviewed by the geopolitical magazine "Geopolitics".txt
 2012-09-26 - AbkhaziaMfa - 92 - Meeting of the Deputy Ministers of Foreign Affairs of the Republic of Abkhazia I.txt
 2012-09-27 - AbkhaziaMfa - 93 - Acting Minister, Deputy Foreign Minister of the Republic of Abkhazia Irakli Khintba.txt
 2012-09-30 - AbkhaziaMfa - 95 - A diplomatic reception on the occasion of the Victory and Independence Day of the Republic of Abkhazia.txt
 2012-10-02 - AbkhaziaMfa - 96 - The Ministry of Foreign Affairs of the Republic of Abkhazia sent a congratulatory letter to the President of the Republic of Georgia.txt
 2012-10-02 - AbkhaziaMfa - 97 - Acting Minister, Deputy Minister of Foreign Affairs of the Republic of Abkhazia.txt
 2012-10-04 - AbkhaziaMfa - 98 - The Ministry of Foreign Affairs of the Republic of Abkhazia sent a congratulatory letter to the President of the Republic of Georgia.txt
 2012-10-08 - AbkhaziaMfa - 99 - Interview of the Deputy Foreign Minister of the Republic of Abkhazia Irakli Khintba.txt
 2012-10-11 - AbkhaziaMfa - 100 - On October 11, 2012 there will be the XXI round of Geneva discussions on security and stability in Georgia.txt

# Why bother?

- “identify widespread patterns of naturally occurring language and rare but telling examples, both of which may be overlooked by a small-scale analysis.”

Baker, Paul, and Tony McEnery. 2005. “A Corpus-Based Approach to Discourses of Refugees and Asylum Seekers in UN and Newspaper Texts.” *Journal of Language & Politics* 4 (2): 197–226.

- Find the needle in the haystack
- Characterise the haystack

Hopkins, Daniel J., and Gary King. 2010. “A Method of Automated Nonparametric Content Analysis for Social Science.” *American Journal of Political Science* 54 (1): 229–47. doi:10.1111/j.1540-5907.2009.00428.x.

# Why bother?

- By treating the internet as an inordinate mass of contents that can be superficially explored through search engines and serendipitous findings, we are missing out:
  - we cannot see trends, we may miss information
- Methodological rigour, replicability, and such

## Some examples

(the point is not “look at what I have done”, but rather “could something like this be useful to answer your own research questions”)

## Examples based on word frequency

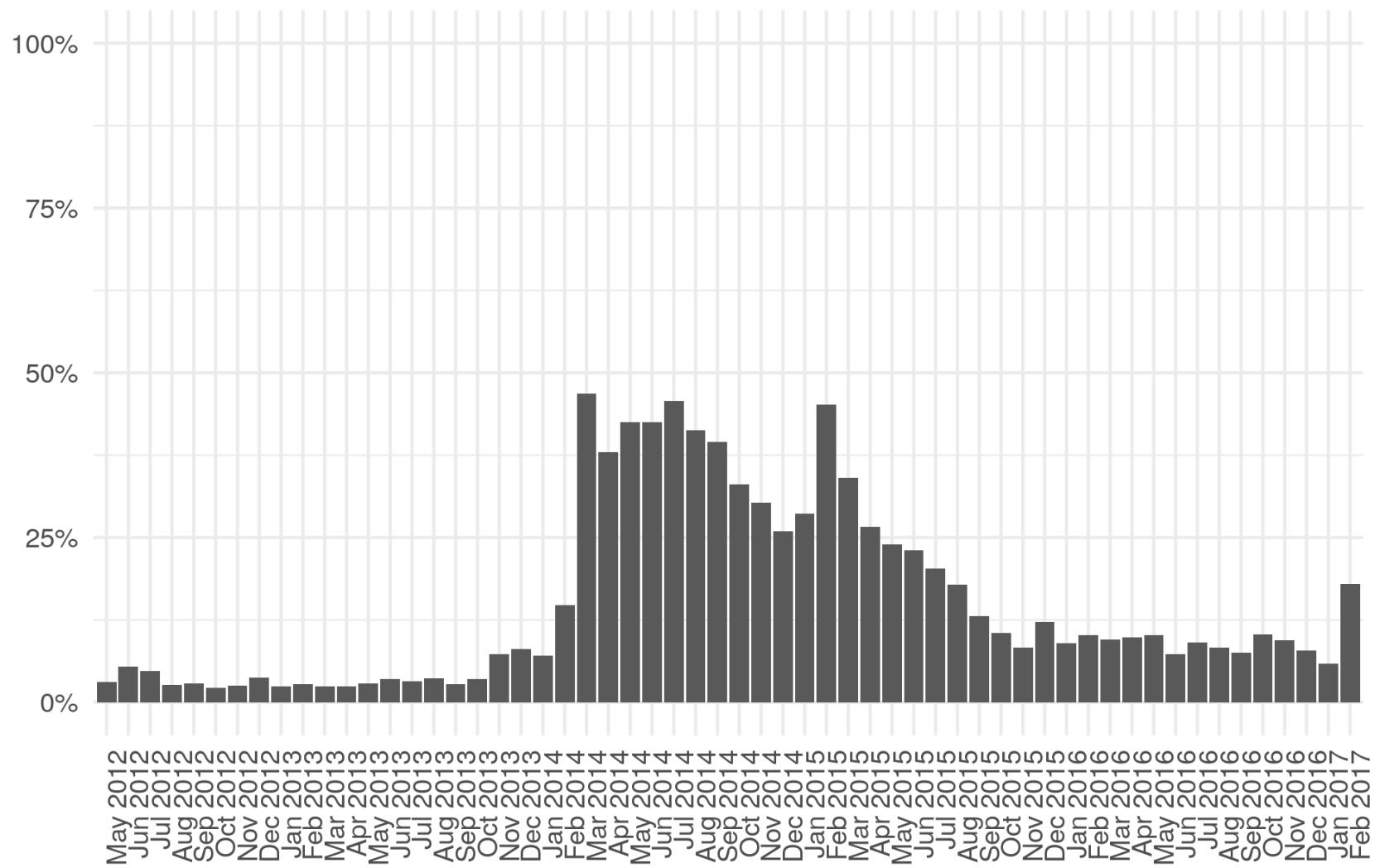
**“The most valuable use of studies of content [...] is in noting trends and changes in content”**

**(Albig, 1938, p. 349)**

Albig, William. 1938. “The Content of Radio Programs, 1925-1935.” Social Forces 16 (3): 338–49. doi:10.2307/2570805.

# Example #1: Russian media and Ukraine

Share of articles including reference to 'Ukraine' on Pervy Kanal

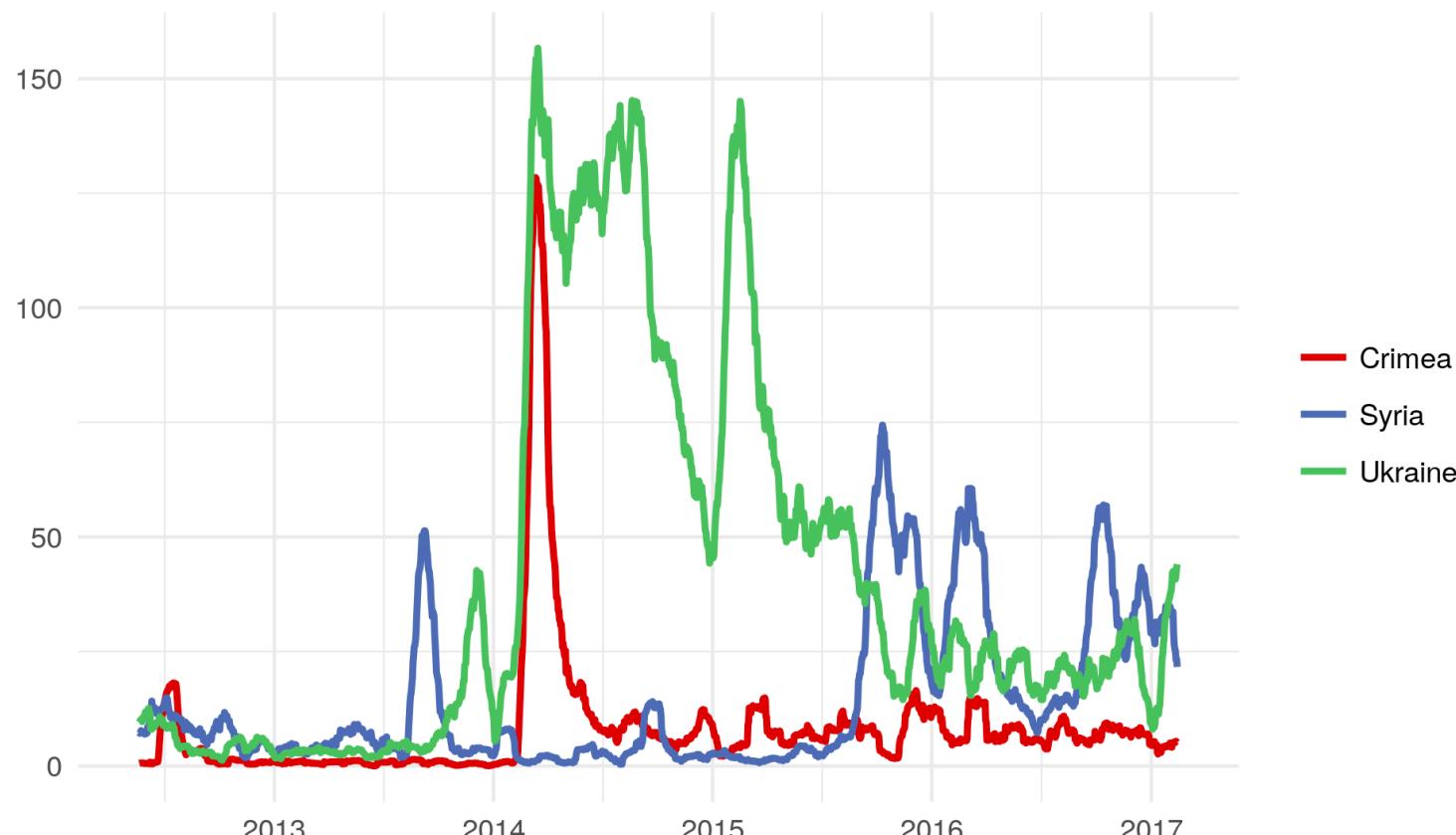


# Example #2: Russian media

## Ukraine/Crimea/Syria

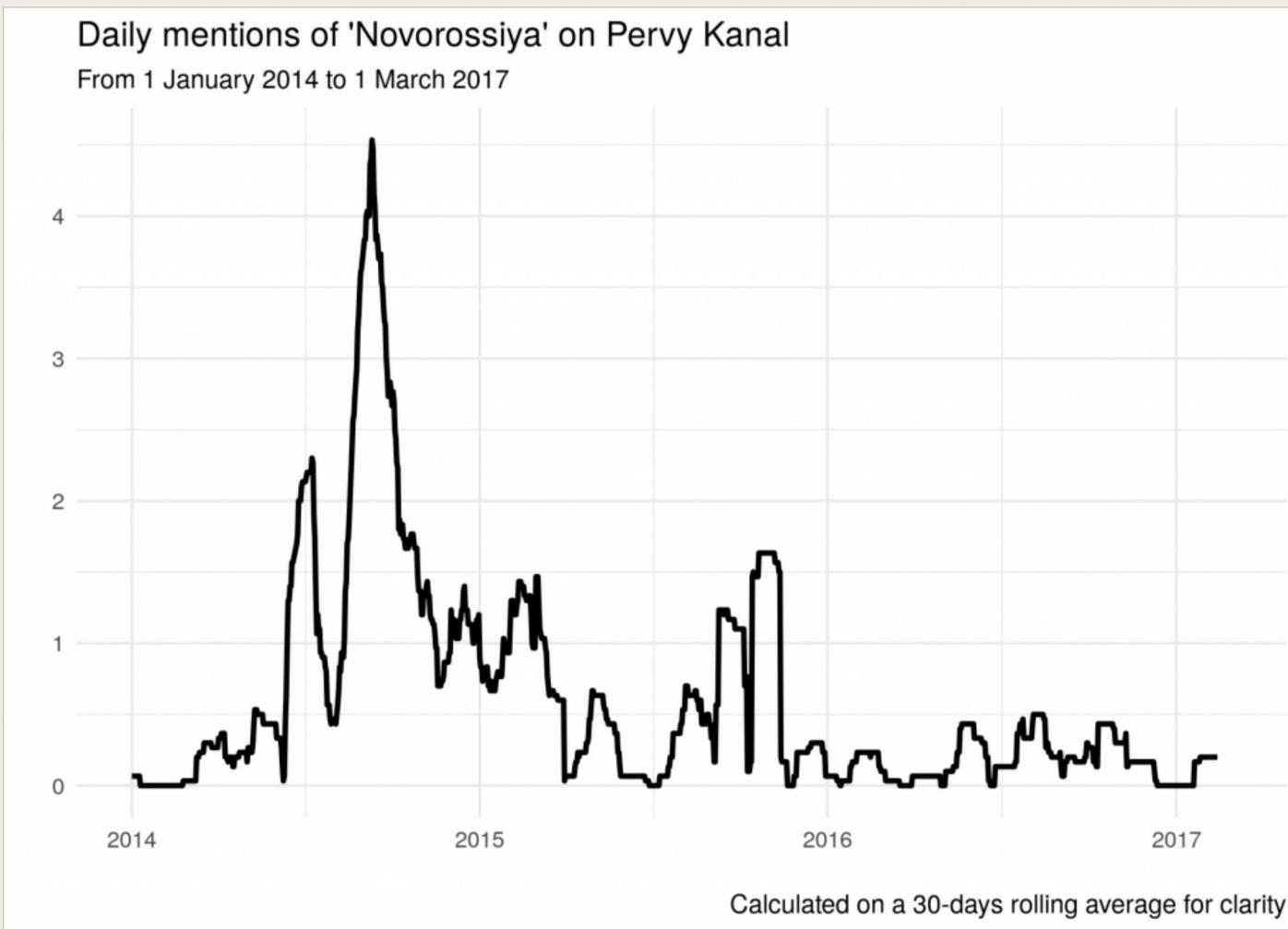
Daily mentions of 'Ukraine', 'Crimea' and 'Syria' on Pervy Kanal

From beginning of Putin's presidency on 7 May 2012 to 1 March 2017



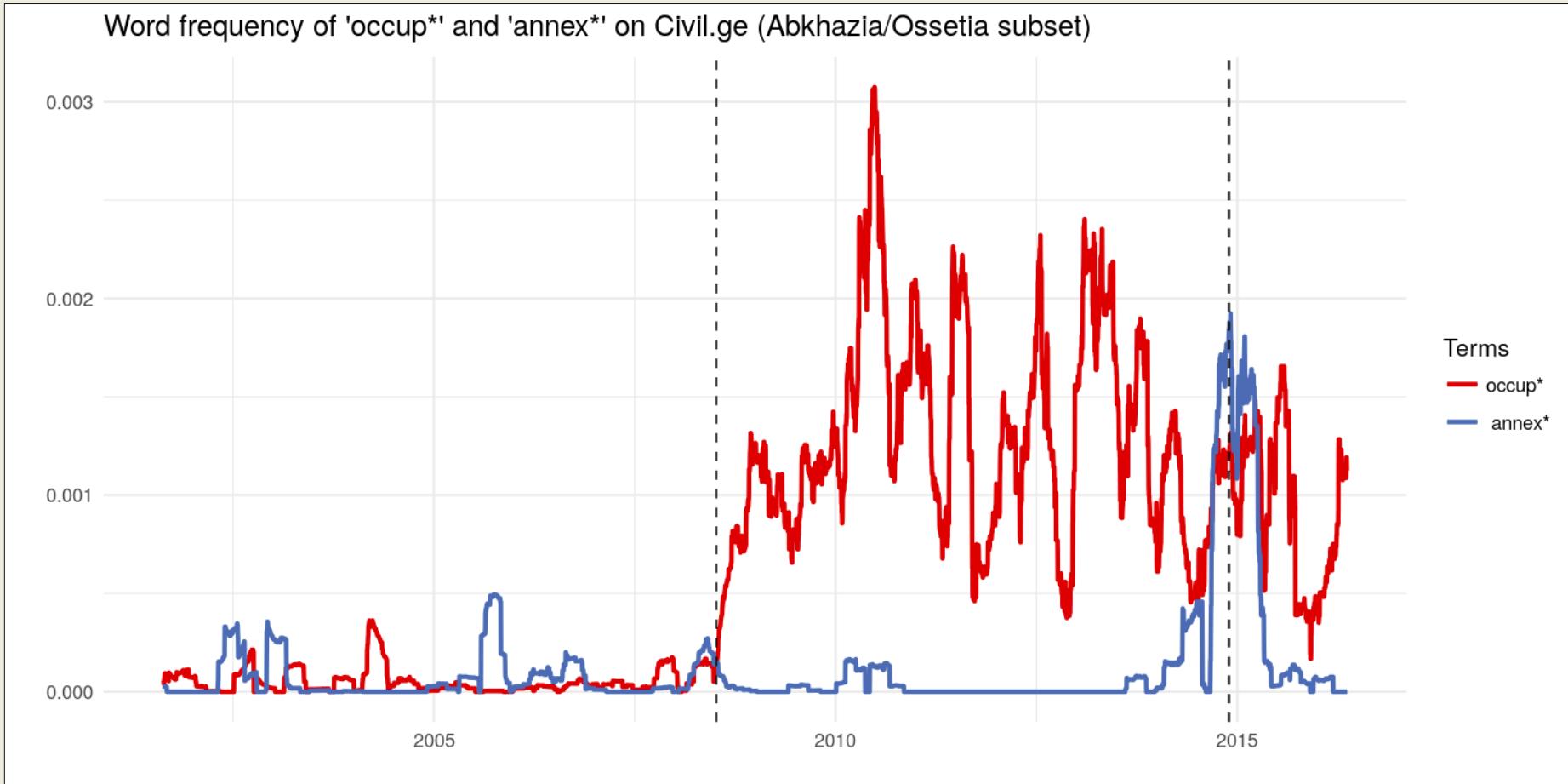
Calculated on a 30-days rolling average for clarity

# Example #3: Novorossiya



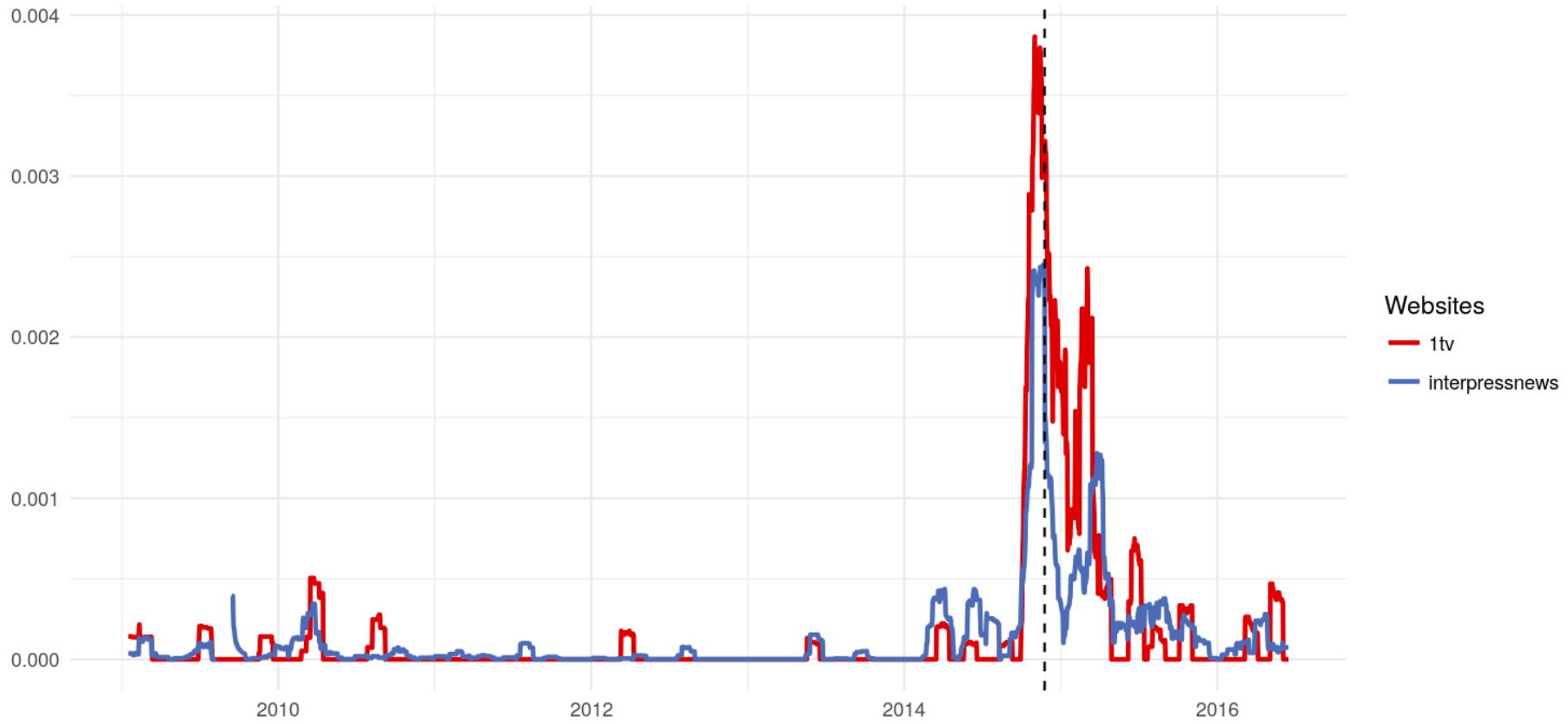
<http://www.giorgiocomai.eu/2017/03/20/word-frequency-of-ukraine-crimea-dnrlnr-and-novorossiya-on-1tv-ru/>

# Local media



# Not necessarily in ‘mainstream’ languages

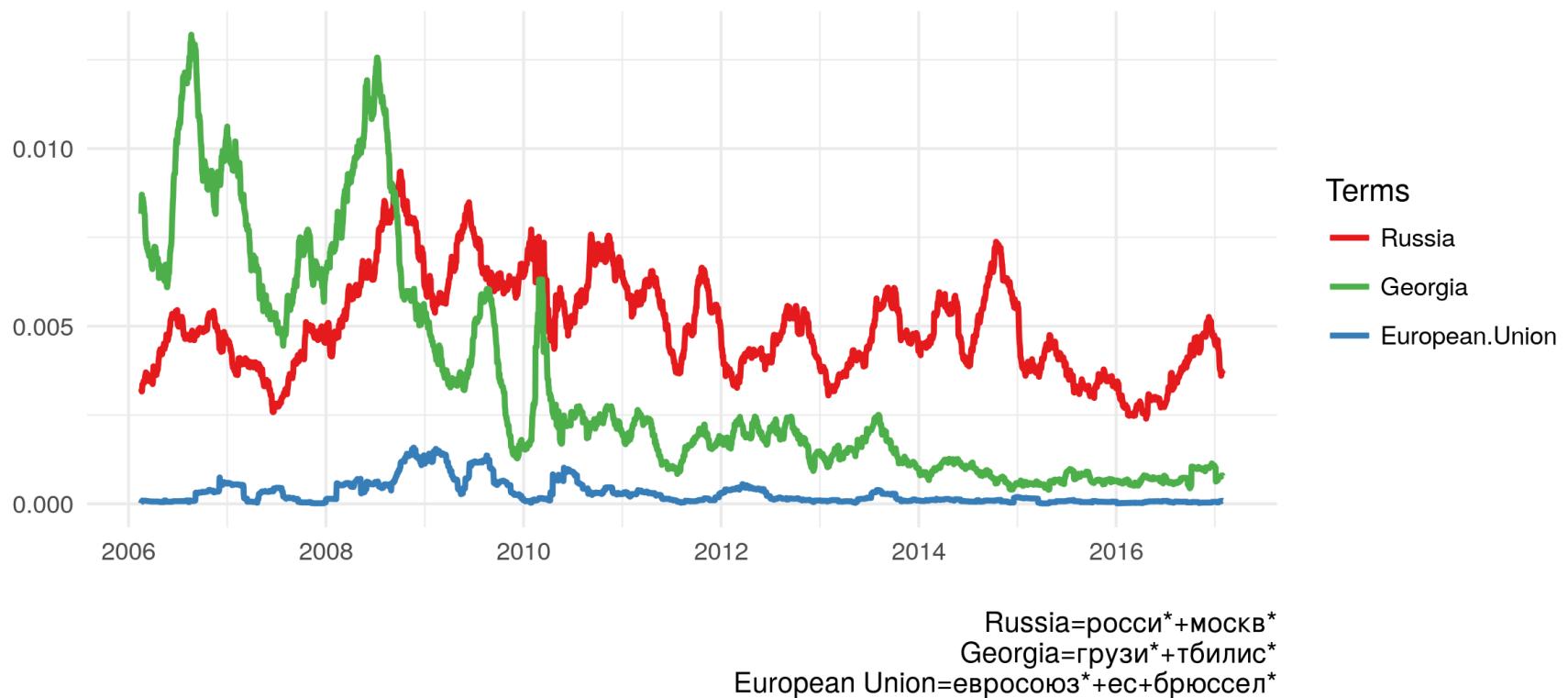
Word frequency of 'annex\*' on 1tv.ge and Interpressnews.ge (Abkhazia/Ossetia subset)



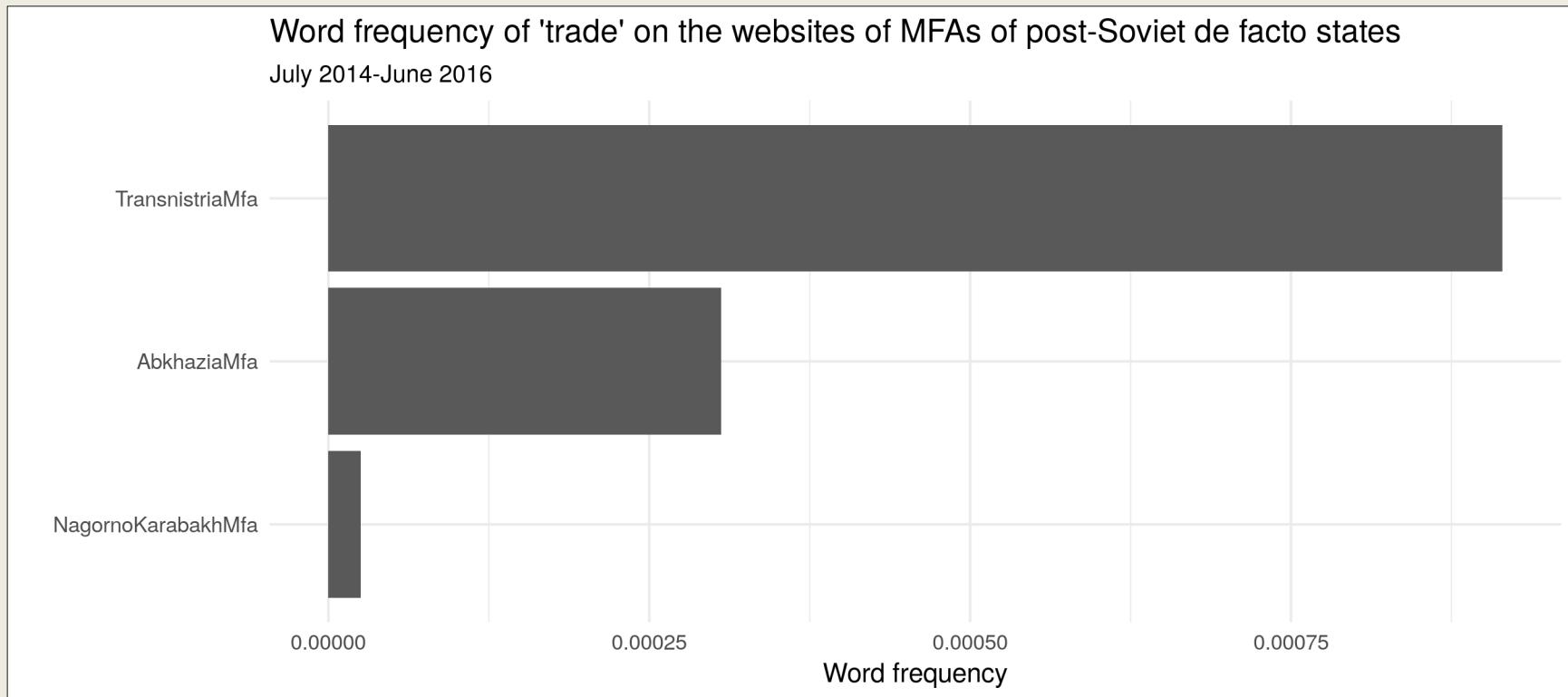
# Or from sources that may not appear in established databases

Word frequency of “Russia”, “Georgia”, “European Union”

Calculated on a rolling average of 90 days for clarity



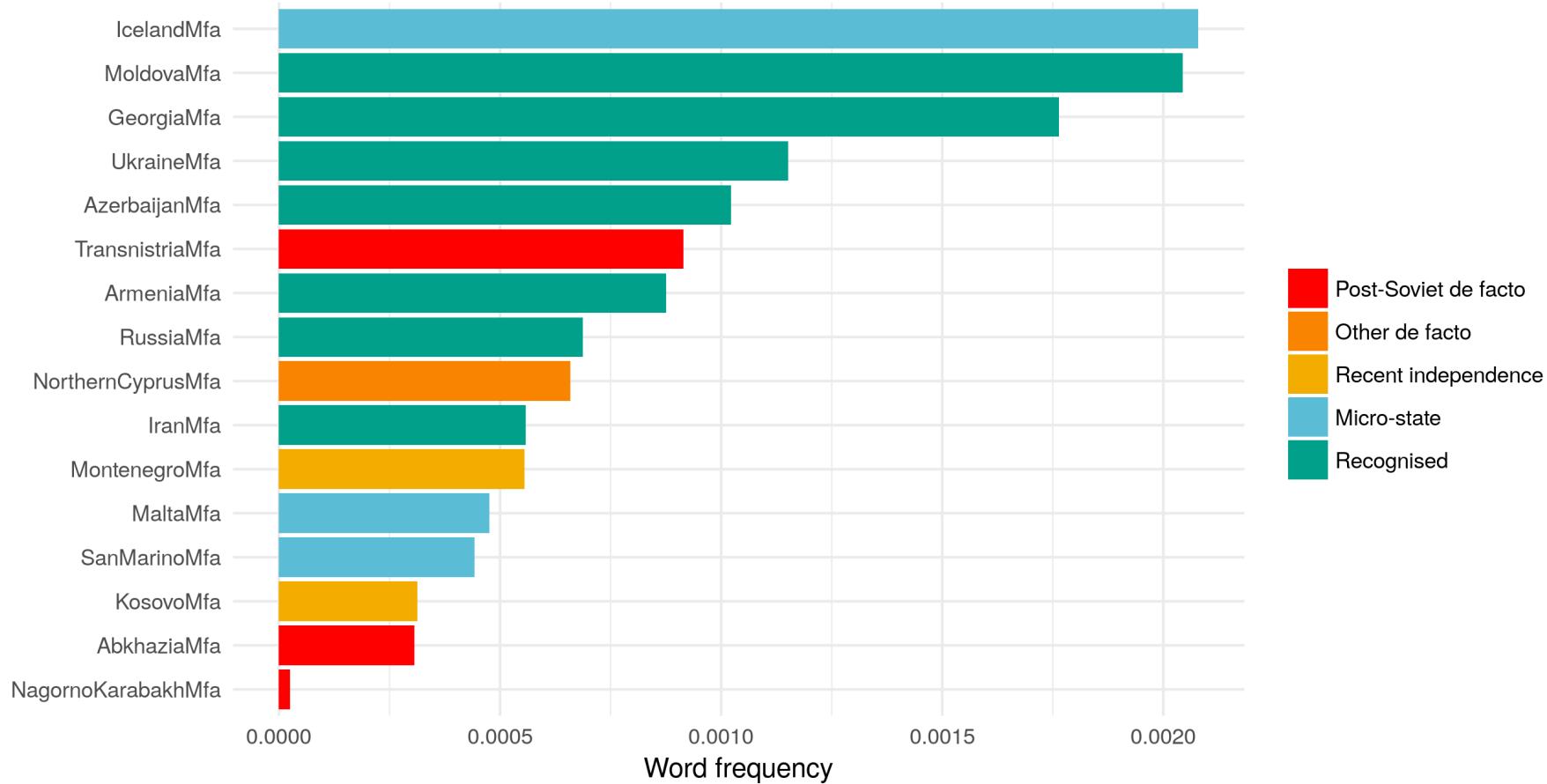
# Such as institutions in de facto states (or local institutions anywhere, really)



# More cases

Word frequency of 'trade' on the websites of selected MFAs

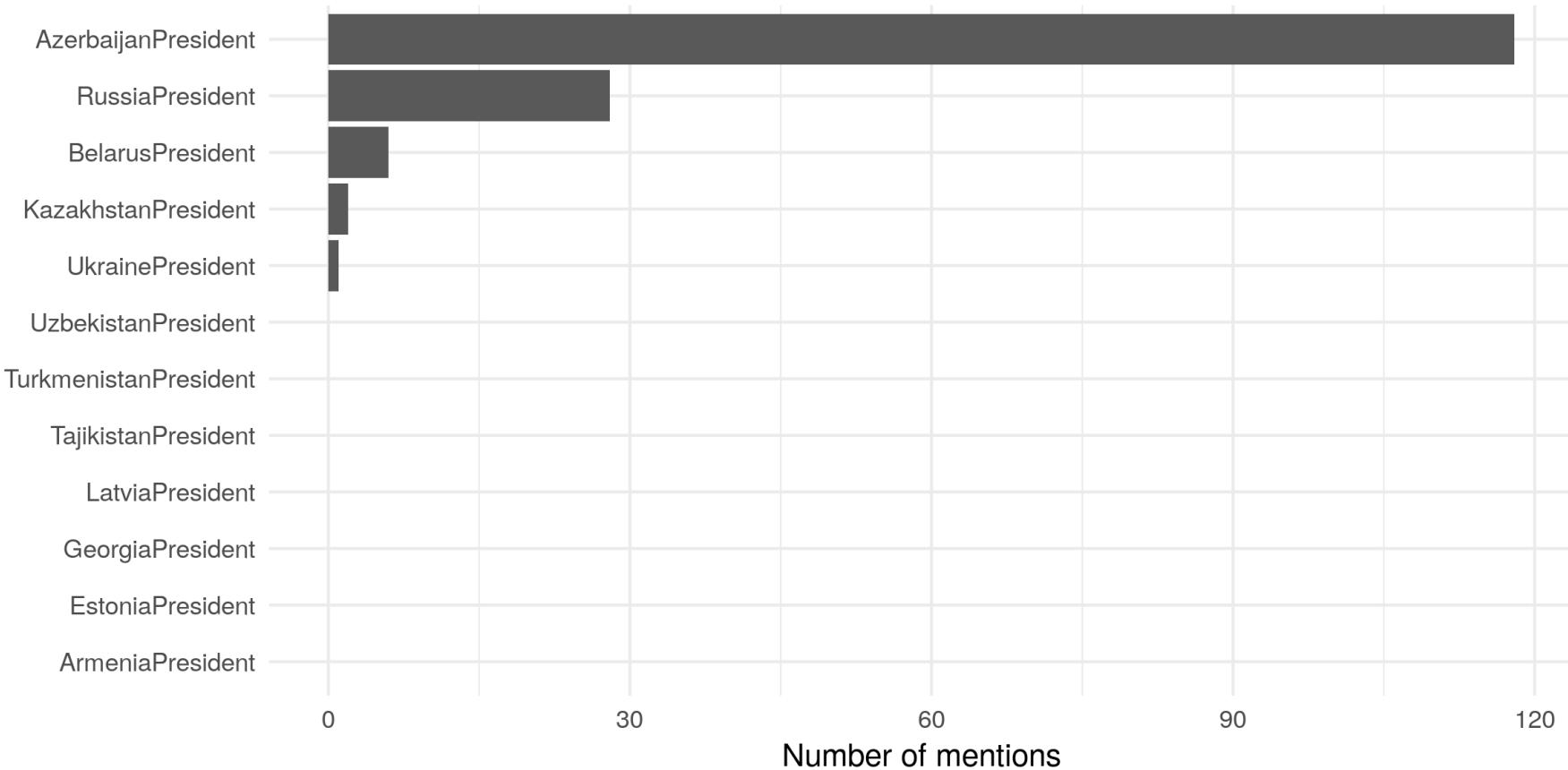
July 2014-June 2016



# To present case selection

Number of references to 'double standards'  
on the websites of presidents of post-Soviet states

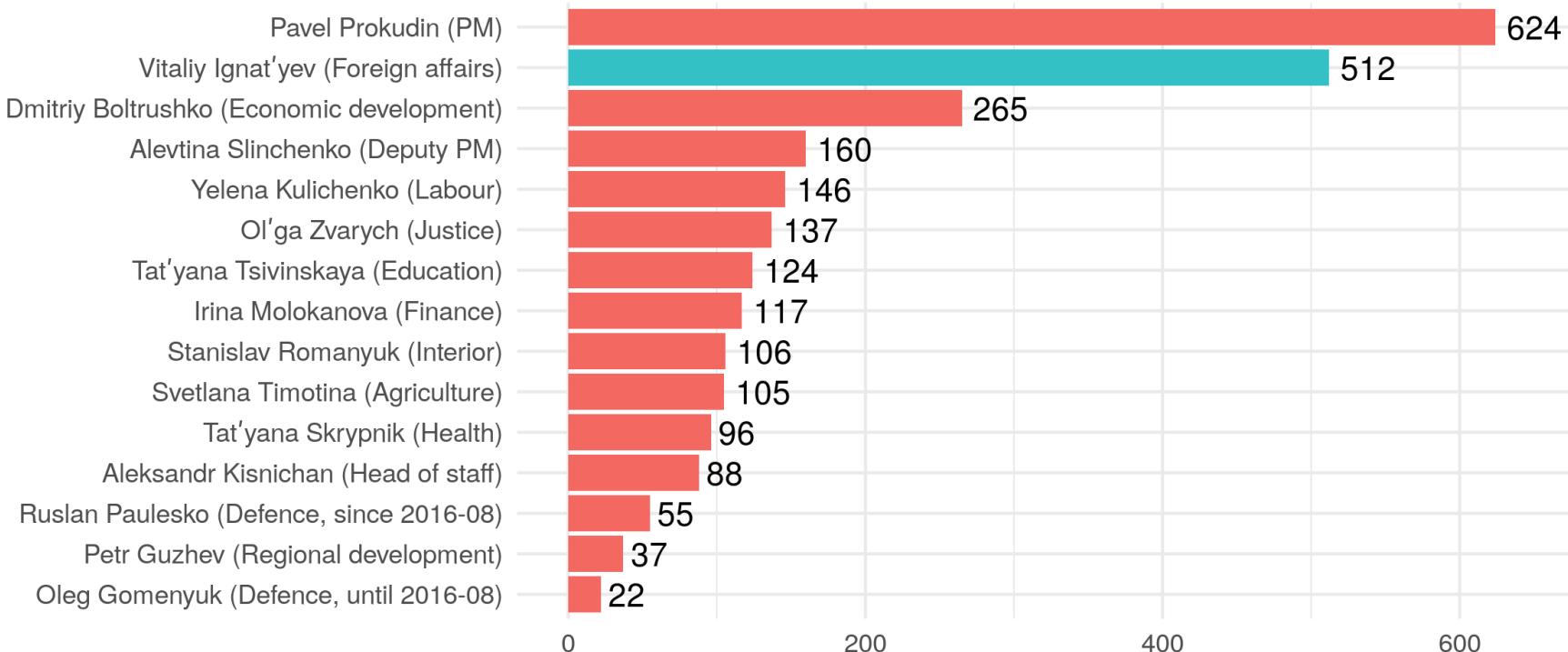
2012-2015



# To point at visibility of politicians, or of the institution they represent

Members of Prokudin's government (Transnistria)

By number of mentions on NovostiPmr.com during the tenure of the government  
(25 December 2015-17 December 2016)



# Or visibility of candidates

Most famous candidates to Abkhazia's parliamentary elections (second round)

By number of mentions on Apsnypress, district, and electoral outcome



<http://www.giorgiocomai.eu/2017/03/31/abkhazias-parliamentary-elections-not-for-the-famous/>

**Or simply to subset materials and analyse  
them qualitatively, find data, etc.**

# Why is this approach so uncommon in area studies?

- (area studies) researchers don't know how to go about it
- It is technically complicated and time-consuming
- Epistemological issues (?)
- ?

Decades after the technique was established, and in spite of technological advancements “content analysis is still an expensive research tool. [...] And it is so even in computer-assisted content analysis. [...] Computer-aided content analysis is still time consuming.” (Franzosi 2008, p. XXXV)

# Are we just lazy?

## Google Books Ngram Viewer

Graph these comma-separated phrases:   case-insensitive

between  and  from the corpus  with smoothing of

[G+ Share](#) 0

[Tweet](#)

[Embed Chart](#)



**If it was easier, would more area-study  
researchers use this approach?**

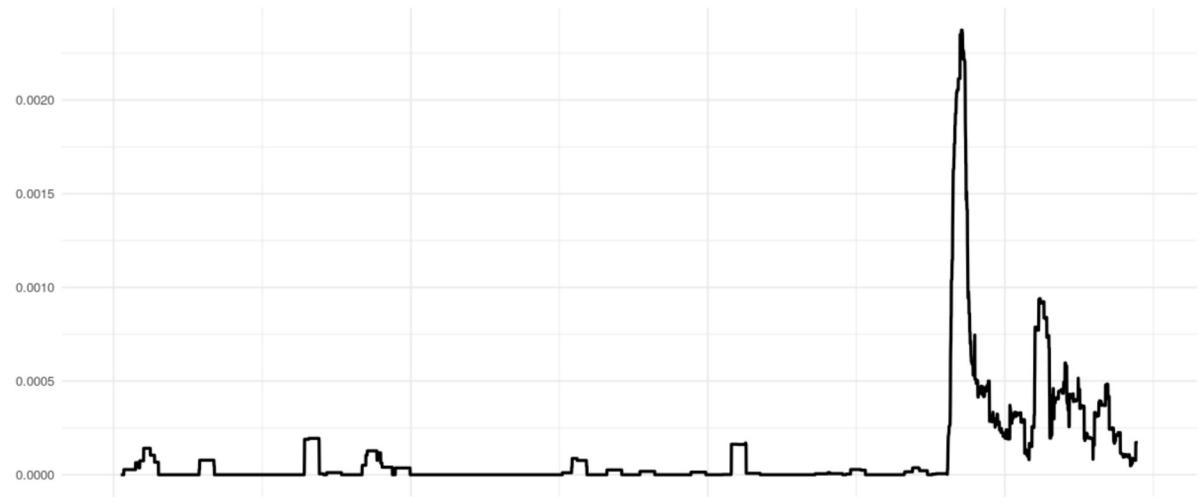
# If it was easier, would more area-study researchers use this approach?

Word frequency on Kremlin.ru

Term to be analysed

crimea

Word frequency of crimea on Kremlin.ru



Calculated on a 91-days rolling average for clarity

# If it was easier, would more area-study researchers use this approach?

## Official websites of South Caucasus presidents

Select type of graph  
 Time series  
 Barchart

Terms to be analysed\*  
europ\*

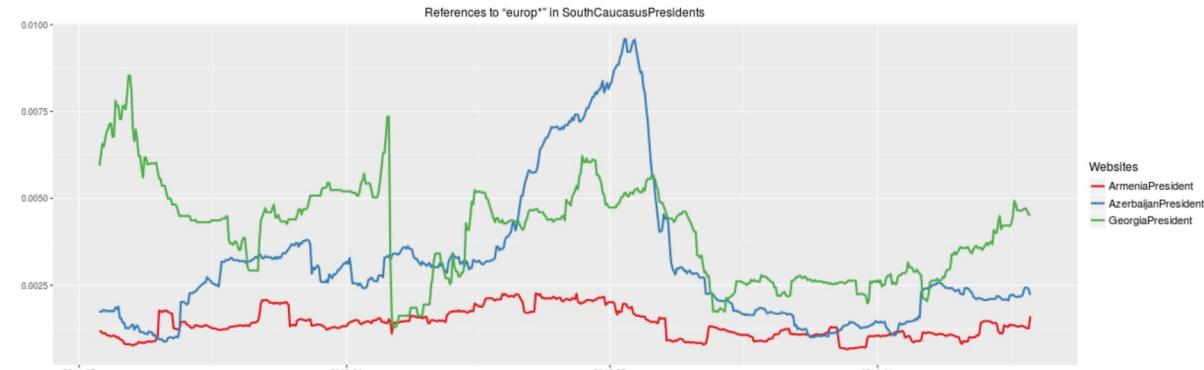
startDate  
2014-06-01

endDate  
2016-05-30

Rolling average (days)  
1  90

Smooth  
 Linear smooth  
 Show interactive time series  
 Show keywords in context

\* If more than one, terms must be comma-separated; to merge multiple terms in calculations, use the format: Russia=Russia+Moscow+Putin, Europe=Europ\*+EU+Brussels



Show 25 entries

Search:

Date	contextPre	keyword	contextPost	Source
2014-06-02	consolidating the collaboration between Armenia and the	European	Union, as well as at further	President Serzh Sargsyan held meeting with leadership of diplomatic service
2014-06-02	reinforcing and advancing the bilateral relations with	European	countries. The President of Armenia drew	President Serzh Sargsyan held meeting with leadership of diplomatic service
2014-06-05	figure skating, the multiple world and	European	champion, the Russian public	President took part in groundbreaking ceremony of new figure skating school and visited companies of Grand Holding

# If it was easier, would more area-study researchers use this approach?

## Official websites of South Caucasus presidents

Select type of graph

- Time series
- Barchart

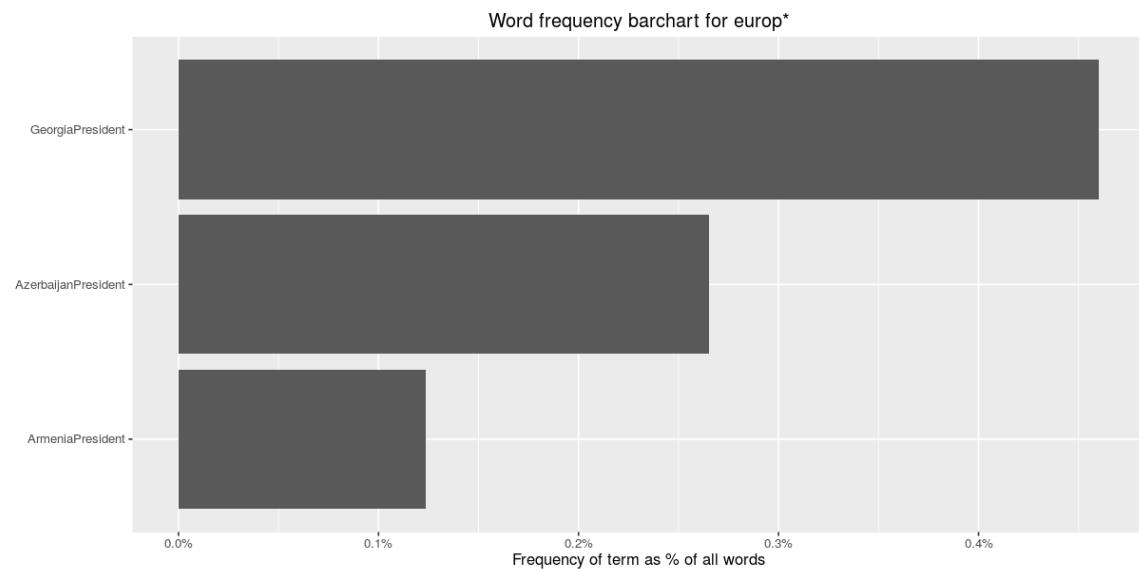
Terms to be analysed\*

Select type of barchart

- Barchart by website
- Barchart by year and website
- Barchart by year (merge multiple sources)
- Barchart (merge multiple sources)

Frequency

- Relative frequency
- Absolute number of occurrences



# Maybe also to share qualitative coding, notes, etc.?

**Total items: 335**

Reset filter

All  Only tagged  Only untagged

**Pattern to be found**

**Country**

**Sector**

**Type**

**Submit**

country: Russia  
sector: sport  
type: NA

Filter

**Pattern to filter**

**Filter by country**

Country

- Abkhazia
- Armenia
- Belarus

**Filter by sector**

Sector

- accounting
- agriculture
- anti-monopoly

**Filter by type**

Type

**Keep items that belong to any of the above categories, or only those that belong to all of those selected?**

Any  All

Invert filter?

**Filter**

Project-website-id: deFactoNews-novostIPmr-37344

## В Правительстве обсудили развитие приднестровского спорта высших достижений

2016-04-18  
<http://novostIPmr.com/ru/news/16-04-18/v-pravitelestve-obsudili-razvitiye-pridnestrovskogo-sporta-vysshih>

Тирасполь, 18 апреля. ИА «Новости Приднестровья». Глава Правительства Павел Прокудин провел межведомственное совещание, в ходе которого обсуждались принципы участия отечественных профессиональных спортсменов в международных соревнованиях в составе сборных признанных государств. На данный момент статус непризнанной республики не позволяет Приднестровью представлять свои команды на турнирах различного уровня. Часто наши спортсмены вынуждены выступать под флагом Молдовы, поднимая своими достижениями престиж молдавского спорта. По мнению Председателя Правительства, такое положение дел наносит вред имиджу нашей страны. «Становится по-настоящему горько, когда ты видишь наших ребят, стоящих на пьедестале, а над ними развивается флаг другого государства», - добавил Премьер. Такого же мнения придерживается и внешнеполитическое ведомство ПМР. Как подчеркнул заместитель министра иностранных дел Дмитрий Паламарчук, Республика Молдова выступает исключительно «потребителем» наших спортсменов. Однако Павел Прокудин отметил, что «без соревновательного процесса никакой прогресс и профессиональный рост невозможен». При этом Председатель Правительства оценил высокий патриотизм наших спортсменов, которые стараются подчеркнуть свою принадлежность Приднестровью. Как сообщает пресс-служба Правительства, отечественные спортсмены успешно выступают в составе российских сборных. Между тем, как отмечали участники встречи, Российская Федерация является не только страной-гарантом, нашим экономическим партнером, но и реализует в Приднестровье социальные проекты, в том числе в области спорта. Меморандум о сотрудничестве, заключенный в феврале 2015 года между Министерством спорта Российской Федерации и Государственной службой по спорту Приднестровья, стал прямым каналом коммуникации. Участники совещания сошлись во мнении, что необходимо переходить к более активному взаимодействию в сфере спорта высших достижений. Также речь шла о повышении квалификации местного тренерского состава и спортсменов в рамках тренингов и семинаров с привлечением ведущих российских специалистов. Отдельное внимание уделили вопросу усиления патриотического воспитания в спортивных школах и секциях. По итогам совещания председатель Правительства дал ряд соответствующих поручений. #Приднестровье#Правительство#спорт

# Concluding remarks

- Actually giving the chance to explore and conduct analysis of the datasets we created (both quantitative and qualitative) opens new possibilities for feedback and alternative explanations
- Creating textual datasets based on websites (media, institutions, etc.) can be easy
- And if we all share them, then it's even easier, and we can build upon each other's work
- If it was easy enough, would researchers do it?

# Text-mining the post-Soviet web

**Giorgio Comai**

*Dublin City University*

*@giocomai*

Comai, Giorgio (forthcoming, 2017). “Quantitative Analysis of Web Contents in Support of Qualitative Research. Examples from the Study of Post-Soviet de Facto States.” *Studies of Transition States and Societies*.

*<http://giorgiocomai.eu/>*

*<https://github.com/giocomai/castarter>  
[code is being rewritten, not fully functional]*

*<https://giocomai.shinyapps.io/kremlin/>*

*<https://giocomai.shinyapps.io/kremlinregex/>*

*<https://giocomai.shinyapps.io/SouthCaucasusPresidents/>*