Quantitative analysis of web content in support of qualitative research: examples from the study of post-Soviet de facto states

Comai, Giorgio

Veröffentlichungsversion / Published Version
Zeitschriftenartikel / journal article

Quantitative Analysis of Web Content in Support of Qualitative Research. Examples from the Study of Post-Soviet De Facto States

Giorgio Comai*

Abstract

In recent years, the internet has been increasingly adopted as a key means of communication by local authorities, organisations and news media throughout the post-Soviet context. This has led to the creation and on-line publication of content that is routinely consulted and quoted by scholars of area studies, who, however, seemingly approach the web as an inordinate mass of content that can be superficially explored thanks to search engines and meaningful keywords. Structured analysis of content is still uncommon in area studies for a few reasons: it is considered to be time consuming, difficult to learn and, fundamentally, relevant datasets are usually not readily available. This paper briefly presents how to overcome these obstacles by introducing an open source package developed by the author that facilitates the creation of structured textual datasets from web content, and allows for basic word frequency analysis in a straightforward web interface.

This article argues in favour of a wider use of quantitative methods based on the analysis of word frequency in textual datasets extracted from the internet as a starting point for in depth research with established qualitative methods. The examples presented in this paper relate to the study of post-Soviet de facto states.

Key words: content analysis, word frequency, de facto states

> "The most valuable use of studies of content [...]
> is in noting trends and changes in content"
> (Albig, 1938, p. 349)

Introduction

In recent years, the internet has become a key source of information for academics working in social sciences and humanities. Even if often not explicitly included among data collection methods, it is considered standard practice to look for relevant information on-line through search engines before doing fieldwork or proceeding with other aspects of research. Such preliminary work, however, is highly dependent on the so-called 'Google skills' of the individual researcher and mostly takes place unsystematically. By its nature, this approach treats the internet as an inordinate mass of content that can be superficially explored thanks to search engines and meaningful keywords.

However, in practice, websites are often highly structured. A research question involving a well-defined territory, institution or community may benefit from a structured analysis of the textual contents of a specific website, a section of a website, or a limited number of websites. Once extracted from the internet, textual content accompanied by metadata (most importantly, date of publication) can be quickly converted into a carefully tailored dataset (or corpus, as it is frequently called in content analysis).[1] This opens the way for quantitative content analysis techniques, as well as the possibility to analyse qualitatively a well-defined subset of materials. As highlighted

* E-mail: giorgio.comai@dcu.ie

1 According to Krippendorf (2004, p. 18), "content analysis is a research technique for making replicable and valid inferences from texts (or other meaningful matter) to the contexts of their use." For an overview of the development of the concept in recent decades, see in particular Franzosi (2008, pp. xxi–xlx).

by Baker and McEnery (2005, p. 198), this allows 'researchers to objectively identify widespread patterns of naturally occurring language and rare but telling examples, both of which may be overlooked by a small-scale analysis.'

This article argues that a structured approach to analysing web content could, and perhaps should, become a common component part of the research process in qualitative research, both in area studies – which are the direct focus of this article – and beyond. Qualitative studies may benefit from a more structured and explicit approach towards analysing on-line contents, an activity that is implicitly included in many studies on current events, recent history or contemporary debates about historical events or figures. Such an approach does not necessarily imply using quantitative methods, and can be used simply to formalise a key component of the research process. Beyond that, basic quantitative analysis of contents based on word frequency can be usefully integrated into qualitative studies, in order to provide additional background information, fine-tune interview guides, or corroborate evidence.[2]

One of the reasons why quantitative content analysis is still relatively uncommon in area studies is that it is considered a technically complex and extremely time-consuming endeavour. As Franzosi (2008, p. XXXV) put it, 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." Indeed, while qualitative content analysis is often based on complex coding procedures and is troubled by issues of inter-coder reliability, quantitative content analysis increasingly involves advanced statistical methods and complex analytical techniques. However, if elaborate content analysis techniques require time, resources and skills that are usually not available to individual researchers, the response must not necessarily be to disregard these methods completely. Instead, a 'back to the basics' approach could be applied, limiting quantitative content analysis to its most basic application: word frequency.

Once the limitations of this method are kept in consideration, and a properly structured dataset has been built, content analysis based on word frequency can still do at a basic level what Albig, one of the forefathers of content analysis, considered to be the 'most valuable use' of the method: 'noting trends and changes in content' (Albig, 1938, p. 349). An additional benefit of this minimalistic approach is that it produces straightforward descriptive statistics and graphs that are meaningful and clear also to informed readers and researchers, who have no specific competence in content analysis techniques or statistical analysis. This allows data gathered with this method to be smoothly integrated with qualitative research and to be widely accepted by an audience – that of area studies scholars – largely unaccustomed to advanced content analysis techniques.

A second reason why content analysis is still uncommon in area studies is that relevant datasets are usually not readily available. This article presents briefly how to overcome this difficulty using a freely available, open source package developed by the author that facilitates the creation of structured textual datasets from web content, and allows for basic word frequency analysis in a straightforward web interface.

Finally, it should be highlighted that this article is focused on methods. Empirical results are not its main focus, but are rather presented 'to demonstrate the methodological point(s)' (Fetters & Freshwater, 2015, p. 208) that are the centre of this paper. All examples are related to the study of post-Soviet de facto states, and the results presented may be relevant for scholars with an interest in these territories.[3] De facto states serve as a particularly illustrative example, since they have long been considered 'informational black holes' (King, 2001, p. 550), local sources are not included in established media databases, and the limited quantitative data available related to them are highly contested.

2 The quantitative and qualitative component of the research may interact in different phases of the research, for example, by 'following a thread', as suggested by Moran-Ellis et al. (2006, p. 54).

3 De facto states (or unrecognised states) can be concisely defined as "entities that have achieved and maintained internal sovereignty over an area for an extended period, with a degree of internal legitimacy but only limited formal recognition at the international level, or none at all" (Ó Beacháin, Comai, & Tsurtsumia-Zurabashvili, 2016).

The next section presents a method to extract textual content from websites with 'castarter', a freely available open source package for the R programming language (R Core Team, 2016) developed by the author. The following parts of the article debate how quantitative analysis of the datasets thus created can be usefully employed as a starting point for in depth research with established qualitative methods, to corroborate information, and, with some limitations, to explore comparison between cases.

Creating datasets

Books dedicated to content analysis typically (e.g. Krippendorff, 2004; Schreier, 2012) do not debate in detail how the dataset is created. They often include sections on sampling (Krippendorff, 2004; Riffe, Lacy, & Fico, 2005), but they still assume that the researcher has already created, has access or might buy access to a structured dataset.[4] Indeed, a significant part of published research applying this method either explicitly makes reference to on-line databases or includes a relatively small amount of content that has presumably been acquired manually (e.g. by copy/pasting individual documents into a database).[5] Commercial databases such as Lexis-Nexis include a number of sources that are relevant for students of area studies, including summaries of local media translated in English by the BBC monitoring service. However, as emerges from the study of de facto states, established databases may prove disappointing when the research question is focused on small and relatively isolated communities, specific institutions, or content available in a relatively uncommon language.

Until a few years ago, independent collection and analysis of a vast number of textual materials from previously uncatalogued official or media sources would have been seriously constrained by the human capacity of the researcher as well as by the difficulty of securing physical access to those materials. The spreading of the internet in much of the world has fundamentally altered this condition. Starting with the early 2000s, and more evidently in the 2010s, the internet has become commonplace also in previously isolated territories; local authorities at all levels throughout the post-Soviet space (including in de facto states) have created their own websites, where they regularly publish new contents.

An overview of websites related to government authorities in post-Soviet de facto states conducted by this author in the summer of 2014 found over 100 websites directly related to government authorities based in these territories, including official websites of local authorities, state departments, customs offices, courts and inspectorates, along with those of key state institutions such as the president, government, parliament and key ministers.[6] Most of them are regularly updated.

Content published on these websites is quoted among sources in academic articles related to de facto states, and the websites themselves have been the object of scholarly analysis. The 'Laboratory for the analysis of the Transnistrian conflict' at the University of Sibiu (Romania), has created a database detailing the main features of a few dozen websites related to Transnistrian organisations and institutions (Laboratorul Pentru Analiza Conflictului Transnistran, 2013). This author provided a brief overview of how post-Soviet de facto states manifest themselves on-line (Comai, 2015a). However, no previously published research has looked at these websites or their contents in a structured way.

In this article, only textual content of websites is considered. Previous studies have drawn inferences by categorising websites according to features such as the languages in which they

4 Riffe, Lacy and Fico (2005, pp. 209–210) suggest a number of existing databases accessible for a fee, such as Lexis-Nexis (http://www.lexis.com/).

5 Kutter and Kantner (2012, p. 6) mention a few such studies.

6 All of these websites are in line with Fursich and Robins's (2002, p. 195) definition of 'government website' - 'Internet sites produced or initiated by a national governmental institution (such as a Ministry of Information), which are sanctioned by the political leadership of the given country' - as well as the criteria brought forward by Mohammed (2004, p. 475), 'To be included in the final data set, sites had to meet three criteria. They had to be: evidently official by statement or content; national in scope; and hosted, sanctioned and or produced by a native government or government agency.'

are available and the name of the web domain (Mohammed, 2004), content categories (McMillan, 2000), or others. However, websites are treated here instrumentally, simply as repositories of textual information, which at another point in time may have been published on local bulletin boards, transmitted directly to newsrooms, read on the radio, or printed on paper. Frequency, features and perhaps even the content of publications may be influenced by the medium of transmission – in this case, the internet – but this is not the focus of this study. Besides, looking exclusively at textual content has a number of advantages, facilitating analysis of mass content and comparison between cases.

Extracting textual content from websites can be a challenging task for scholars less familiar with programming or with the inner workings of websites. The data shown in the following sections of this article have been collected using Castarter – Content Analysis Starter Toolkit for R,[7] a package originally developed by this author for the open source R software (R Core Team, 2016), based on previous work in particular by Feinerer, Hornik, and Meyer (2008), Wickham (2009), Bouchet-Valat and Bastin (2013), Benoit and Nulty (2016), and Silge and Robinson (2016). Castarter facilitates the process of finding direct links to individual pages of specific sections of a website, retrieving them, extracting core textual content and key metadata such as date and title (see Illustration 1). The resulting output is stored in an internal database and can be exported as a folder, as text files or as a spreadsheet that can easily be imported in other software packages for further analysis. Castarter also facilitates the creation of a straightforward web interface that allows for basic word frequency analysis of newly created datasets. This interface can then be accessed on a local computer or shared as a web link, enabling access to other users directly from their browser, without requiring to install any additional software.

Depending on the research question and the kind of analysis that follows, standard procedures such as removing stopwords (frequent and usually irrelevant words such as 'and', 'or', 'while', etc.), applying stemming (reducing words to their stems so that 'example' and 'examples' are identified as one and the same word), removing sparse terms (terms found only in a marginal number of texts), or combining and substituting words (so that, for example, 'European Union' and 'EU' can be counted together), may be applied when appropriate. Such procedures have a substantial impact on the results of the analysis and must, therefore, be applied and verified with caution.

With Castarter, downloading all of the textual contents of a section of a website and extracting relevant metadata may require only a few minutes of a researcher's time in a best-case scenario, and in most cases not more than one hour, including basic operations for finalising and cleaning up the dataset.[8] More details on how to create, analyse and export datasets with Castarter can be found on the project's web page. The following sections present use-case scenarios for datasets created with this method, providing examples related to the study of post-Soviet de facto states.

The footnote you wished you had

Researchers who are thoroughly familiar with a topic believe (and often, rightly so) that they 'know things', even if they would find it difficult to provide specific evidence for some of their claims. For example, anybody familiar with Georgia's political discourse and government rhetoric may point out that in recent years Abkhazia and South Ossetia have been routinely described as 'occupied territories', a characterisation that would have been uncommon in the early years 2000s. Indeed, this change of perspective has been highlighted in the literature, for example, by O'Loughlin, Kolossov, and Toal (2011, p. 4), who pointed at a "powerful tendency […] [that] led in the wake of the August 2008 war to the re-emphasis by Tbilisi of South Ossetia and Abkhazia as 'occupied territories'." This terminology has been enshrined in legislation with the October 2008 Law of Georgia on Occupied Territories (Law of Georgia on Occupied Territories, 2008).

7 Castarter – Content Analysis Starter Toolkit for R – is an open source project under active development. Install instructions, as well as a brief tutorial, are available on the project's page: https://github.com/giocomai/castarter.

8 This estimate does not include the time that the computer needs to actually download all of the pages; depending on the number of articles, this may take a few minutes or many hours. However, this is a completely automatic process that does not require interaction from the side of the researcher.

The claim that there was a change in terminology is the result of the authors' own observations: it is presented and explained very clearly, but no piece of evidence is provided to support their claim. Such statements are common in academic publications and they are often inevitable: it is not always possible (or perhaps, even desirable) to provide detailed evidence for each piece of information mentioned in an article. However, if there was a relatively easy and straightforward way to provide some evidence that this statement is accurate, for example, with an unobtrusive footnote, authors would do well to include it in their writing. Looking for such supporting evidence may also lead to more accurate statements, for example, pointing at the moment when this change of rhetoric actually took place: did it happen with the coming to power of Mikhail Saakashvili in Georgia? With the war in August 2008? With Russia's recognition in late August 2008? Or only after the Law on Occupied Territories was approved in October of the same year?

It is argued here that when relevant evidence can be extracted from freely available on-line resources, researchers should make the effort to provide it. In this case, for example, a useful source of information may be Civil.ge, a respected English language news website that focuses solely on Georgia and routinely includes statements by the president, members and supporters of the government, as well as opposition representatives and experts. All of its content since 2001 is freely available on-line. If there was a substantial change in terminology in the public debate around these territories, this should become apparent with a basic content analysis based on word frequency of Civil.ge.

Indeed, as appears clearly from Illustration 2 and Illustration 3, references to occupation in articles on Civil.ge that mentioned either Abkhazia or Ossetia[9] became much more common starting with 2008. The change is evident: even if references to 'occupation' were not unheard of in previous years, they became commonplace starting with 2008, and more precisely in the aftermath of the war.

Did other words related to these territories enter the political vocabulary in recent years, besides 'occupation'? What about 'annexation'? Once the dataset is created, this becomes a straightforward question. Illustration 4 is based again on Civil.ge and it shows that 'annexation' became more frequently used only in Autumn of 2014, and then only for a brief period. References to 'annexation' in late 2014 in this subset are mostly related to the 'strategic partnership' treaty between Abkhazia and Russia signed on 24 November 2014.

The fact that this apparent change in rhetoric happened more recently, however, offers the possibility to verify the existence of this trend also in other sources that were not available for earlier years, including Georgian language ones. For example, both the current website of Georgia's public broadcaster (1tv.ge) and one of the most established news websites in Georgia (interpressnews.ge) give free access to all news items posted since late 2008. It is now possible to replicate the same operations conducted for Civil.ge: create a subset including only those news items that make reference to either Abkhazia or Ossetia, and draw a timeline of the word frequency of 'annexation'.[10] The resulting graph (Illustration 5) confirms the observations that emerged from analysing Civil.ge, i.e. that references to 'annexation' have been relatively uncommon in recent years and became frequent for a short period in late 2014 at the time of the 'strategic partnership' treaty between Abkhazia and Russia.

To conclude, O'Loughlin, Kolossov, and Toal (2011) characterised accurately the increased use of 'occupation' in the Georgian public discourse in reference to Abkhazia and South Ossetia. However, authors may not always be so confident about their observations; they may want to back up their claim with credible evidence, and, fundamentally, to verify its accuracy or add more detail. If increased availability of structured textual content on-line offers researchers a practical, unobtrusive way to find and present additional evidence, leading to more accurate and stronger

9 For the aim of this analysis, all derivative words of a term are counted as if they actually were one and the same word, for example occurrences of 'occupied' have been counted as 'occupation', and occurrences of 'Abkhazian' have been counted as 'Abkhazia'. Out of a total of 27,447 articles published on the English language version of Civil.ge before the end of June 2016, 8,671 mention either Abkhazia or Ossetia.

10 References to Tskhinvali are here considered the equivalent as references to South Ossetia, since the official name for South Ossetia in Georgia is Tskhinvali region.

substantive arguments, they should welcome it in their repertoire. It should also be highlighted that such evidence may be briefly presented in a footnote, should not be a distraction from the main substantive arguments being presented, and does not necessarily lead to substantial changes in the wider research design.

Case or key moments selection

Due to their lack of international recognition, post-Soviet de facto states have limited formal interactions with the outside world. This is particularly true for representatives of the Ministry of Foreign Affairs (MFA); as Ker-Lindsay (2015, p. 278) put it, "[t]he one official post that presents a problem in almost all cases is the foreign minister of a contested state." Given the fact that any such meeting can be presented as proof of the international standing of the local government and some sort of recognition (or acceptance) at the international level, whenever a representative of the MFA of a de facto state has a meeting involving a foreign official, this is duly reported in a press release on the MFA's official website.

A quantitative content analysis of the news section of the official website of the MFA of a de facto state should thus provide information about which countries are more frequently involved in such interactions or when relations with a given country started to take place. We would expect frequent mentions of the patron state (reporting meetings, cooperation, and agreements), of the parent state (mostly, to denounce its policies or to voice past grievances), of other states or de facto states that have recognised the independence of the territory (recording meetings or publicising congratulatory statements), and perhaps countries where significant communities of ethnic kins are located. Beyond these, which countries are more frequently mentioned in press releases of a given de facto state? Are there active relations with EU countries? Which of them receives more attention? Such information may not be self-evident and could be useful for defining more effectively the focus of research.

Looking, for example, at the news published on the website of Abkhazia's MFA, a simple word frequency bar chart (see Illustration 6) shows that many of the countries mentioned most frequently can easily be ascribed to one of the above-mentioned categories. Besides them, most frequently mentioned countries include Italy, Ukraine, China and San Marino. There would be no a priori reason to expect any of them in the top spots. Interestingly, besides Italy none of the large EU countries is mentioned to any meaningful extent or with any regularity. A quick overview of a subset of the corpus mentioning these countries makes clear that Ukraine is mentioned mostly with a negative connotation, in relation to events that took place there in 2014; there are no signs of cooperation or formal relations. China is mentioned in relation to a series of visits and meetings organized by the 'representative of the MFA of Abkhazia in the People's Republic of China Ge Zhili.' A quick look at the correspondent time series shows that there are not regular contacts with China. Mentions of San Marino are similarly sporadic. Italy, however, is mentioned much more regularly, in relation to a diverse range of meetings and interactions, including meetings with representatives of local government of Italy and trade chambers that led to the joint signature of formal agreements.

A researcher interested in Abkhazia's efforts to strengthen ties with European countries may now consider appropriate to reframe the question in the following terms: why does Abkhazia's government entertain regular formal interactions with actors based in Italy, but not in other EU countries? At this stage, the researcher can use the knowledge acquired through this quick exercise to proceed with their research using traditional qualitative methods. On the one hand, this approach allows to demonstrate that there is no bias on the side of the researcher if they decide to focus in particular on Italy.[11] On the other, it makes it easier to do the 'homework' before fieldwork and interviews, allowing the researcher to go quickly through each official interaction between Abkhazia's MFA and Italian actors reported on the MFA's website.

11 Given the fact that this author is originally from Italy, a particular focus on this country may easily raise the suspicion of bias.

If, however, a researcher is interested in the broader context of interactions between Abkhazia and EU countries, looking at local media, rather than exclusively at press releases issued by the local MFA, may offer additional insights. In addition, creating a textual dataset of content available on ApsnyPress, the news agency run by Abkhazia's de facto authorities, gives the possibility to include a longitudinal dimension, since content is available on its website starting with 2006.

A time series presenting the word frequency of the parent state (Georgia), the patron state (Russia), and the European Union, provides an immediate picture of the relative importance of these actors in the local public discourse, and how their prominence changed through time (see Illustration 7). Before the war in 2008, Georgia was mentioned more frequently than Russia, but in recent years it has been included less and less frequently in media reports, highlighting how Georgian issues and Tbilisi's initiatives are becoming increasingly irrelevant to Abkhazia's public debates. Significantly, references to the European Union are barely visible on this graph. Adjusting the scale and focusing specifically on the EU, however, allows to highlight ongoing dynamics, and namely the fact that in the aftermath of the 2008 conflict, there was a boost of attention towards the European Union, but that year after year the number of references to the EU decreased (see Illustration 8). As it appears from an overview of relevant media reports, this reflects criticism of the European Union's policy of non-recognition as well as of its monitoring mission established in 2008 (EUMM), mixed with comments and demands demonstrating perhaps some hope that the EU would eventually acknowledge new developments and change its approach towards Abkhazia. With time, local public figures stopped appealing to the European Union and EU member states for recognition and engagement, and increasingly ignored (rather than publicly condemned) EU's policy of non-recognition, thus leading to a noticeable decrease in the frequency of mentions.

Interestingly, the graph highlights that 2012 presents an exception to this trend, with a relatively high frequency of references to the European Union, apparently due to efforts by newly nominated EU's special representative in the South Caucasus Philippe Lefort to give new life to EU's declared policy of "engagement without recognition", as well as to visits of diplomats of EU member states to Sukhumi. An overview of relevant articles clearly presents the dominating attitude of Abkhazia's authorities towards such renewed efforts, as appears from statements made by de facto president Aleksander Ankvab in occasion of interactions with diplomats from EU member states: "If they want to cooperate with Abkhazia, then they have to change their attitude towards us and take example form Russia, which gives practical help and support" (26 June 2012); "I am telling you openly that by interacting less with European diplomats accredited in Tbilisi, we in Abkhazia will not be worse off" (11 September 2012); "[Abkhazia's] president said once again that he does not see the meaning of further meetings with representatives of EU countries" (21 February 2012).[12]

Combining time series graphs based on word frequency with a few quotes extracted from key moments allows to provide an effective and concise overview of the context in which a given dynamic is taking place (in this case, relations between Abkhazia and EU countries). This outcome is important in itself, and provides a useful introduction to further research with established methods. In addition, and crucially for the researcher's workflow, the time series allow to spot important turning points or long-terms dynamics that may inform research or prompt new questions, but could otherwise be easily overlooked. Besides, sub-setting the textual dataset by keyword and focusing on specific periods makes it possible to quickly skim through a manageable number of materials.[13]

In some cases, graphs such as the ones included in this section may not be included in the final research output. Instead, they are part of preliminary work and as such they should provide additional information, leading to a better understanding of the issues being researched, limiting

12 Apsnypress.info, "Aleksandr Ankvab: tem, kto khochet sotrudnichat' s Abkhaziei, sleduet brat' primer s Rossii", 26 June 2012, http://www.apsnypress.info/news/aleksandr-ankvab-tem-kto-khochet-sotrudnichat-s-abkhaziey-sleduet-brat-primer-s-rossii-/; Apsnypress.info, "Aleksandr Ankvab prinyal posla Niderlandov v Armenii i Gruzii", 11 September 2012, http://www.apsnypress.info/news/aleksandr-ankvab-prinyal-posla-niderlandov-v-armenii-i-gruzii-/; Apsnypress.info, "Aleksandr Ankvab ne pitaet illyuzii po povody uluchsheniya otnoshenii s Evrosoyuzom", 21 February 2012, http://www.apsnypress.info/news/aleksandr-ankvab-ne-pitaet-illyuziy-po-povodu-uluchsheniya-otnosheniy-s-evrosoyuzom-/ .

13 This is what some researchers are already doing by other means, e.g. by using search engines or the internal search function of a given website. However, the results are not always comprehensive and rarely provided in a systematic form.

researcher's bias, and, for example, more targeted interview questions. Finally, this approach allows the researcher to explore on-line content in a structured and replicable way that can be briefly described in the methodological section of an article, thus bringing into the open what is often a 'hidden' part of the research process.

Comparison between cases

The approach outlined in this paper may be employed to facilitate comparison between cases. At the most basic level, it can be useful to illustrate key differences between a small number of cases. For example, given its economic structure and its geographical location, trade-related issues are a top priority for Transnistria's MFA, much more so than for the MFAs of other de facto states. To anybody familiar with post-Soviet de facto states, Illustration 9 – highlighting Transnistria's MFA frequent references to trade issues – is not going to be surprising. Yet, such a graph may still be helpful in illustrating a point to readers less familiar with these territories.
Given the relative simplicity of adding new cases to the dataset, however, it may be worthwhile trying to extend the comparison to a wider set of cases in order to understand how 'normal' are the MFAs of post-Soviet de facto states. For example, is it common for a Ministry of Foreign Affairs to dedicate this level of attention to trade? How much do press releases of de facto states differ from MFAs of internationally recognised countries in the region, or of similarly sized territories elsewhere in Europe?

A quantitative content analysis of the press-releases of the MFAs of a relevant set of countries may offer a useful starting point to understand some key aspects in which they differ from recognised countries. With this aim, MFAs of post-Soviet de facto states will be compared with the MFAs of parent states (Moldova, Georgia, Azerbaijan), countries bordering with the de facto states (Ukraine, Russia, Armenia, Iran), de facto states elsewhere in Europe (Northern Cyprus), countries of recent independence in Europe (Kosovo and Montenegro), as well as similarly sized micro-states in Europe (Malta, Iceland and San Marino).[14]

For this example, in order to offer consistent results, only the English language version of the websites of MFAs will be considered; English is not an official language of either of the territories included (with the exception of Malta), and it is assumed here that a self-respecting MFA in Europe today would use English to highlight its international activities, and would write about them on its official website. The analysis will be limited to the two-year period between 1 July 2013 and 30 June 2016, for which data is available for all the MFAs included in the comparison.[15]

Preliminarily, it may be interesting to look at how frequently all of these MFAs publish. A quick overview (Illustration 10) shows that MFAs of post-Soviet de facto states mostly publish less often than the bigger and recognised states that border them, but more often than similarly sized but wealthier and long-established micro-states elsewhere in Europe, such as Iceland and San Marino, which may have less pressing concerns. Even if this cannot be considered a meaningful proxy of the level of activities of these MFAs, issuing regular press releases in English should not be taken as a given. Indeed, the MFA of a recognised country such as Moldova has not been able to publish regular press releases in English, and sometimes has left the English language version of its website without updates for entire months.[16]

Looking at the contents of these materials offers more opportunities for comparison. Testing for the word frequency of 'trade' (Illustration 11), it is easy to notice that Transnistria is not an outlier anymore, and that it is rather the other MFAs of post-Soviet de facto states that refer to 'trade' unusually rarely for an MFA.

14 The MFAs of smaller countries such as Andorra, Liechtenstein and Monaco are not included because they do not have a full-fledged English language website. A structured debate of the case selection goes beyond the scope of this article.

15 Table 1 details the date of earliest available publication for each of the websites included.

16 For example, an archived version of the home page of Moldova's MFA retrieved in July 2013, still shows only news from 2012 (Moldova's MFA 2013). Since 2015, the website of Moldova's MFA has been working thanks to support of foreign donors and carries a notice on its home page that is more characteristic of a civil society project, rather than of an MFA of a sovereign state: "This web site was developed with the financial support of the Estonian Government within the Project "Building Institutional Capacity of the Ministry of Foreign Affairs and European Integration" implemented by the United Nations Development Programme (UNDP), Moldova. The opinions expressed in this website are the authors' opinions and do not necessarily reflect the views of the Estonian Government, UNDP Moldova."

MFAs of post-Soviet de facto states may be expected to mention unusually frequently words related to their claims to sovereignty (such as 'independence' and 'recognition'), or to have a disproportionate amount of press-releases dedicated to ceremonial aspects (such as celebrating anniversaries), given the relative scarcity of official interactions worth mentioning in which they are involved. The results (see Illustration 12, Illustration 13, and Illustration 14) are mostly in line with expectations, with the exception of Transnistria, which leads to the tentative observation, to be investigated with other methods, that Transnistria's MFA is more focused on pragmatic activities than the MFAs of other post-Soviet de facto states.

As highlighted in the introduction to this article, these graphs are not presented here for their intrinsic value, but rather as an example of the potential opportunities offered by this approach. If datasets including a dozen or more cases from different institutions or media can be generated rather easily, then scholars may well consider including such analyses at different stages of their research. Even limiting the analysis to basic word frequency, multiple use cases are possible: for example, a subset of the data focusing on a short period of time could be used to compare reactions to an event, or time series could be used instead of bar charts to highlight trends. Besides, such datasets can be shared, used and integrated by other researchers for potentially unrelated research; for example, datasets of a relatively large number of MFAs across the region such as the one presented in this paper could be integrated with other cases and used for research focused on different issues, from state identity to foreign policy priorities.

Results obtained with primitive tools such as the ones demonstrated above should not be considered exhaustive evidence, but rather as additional information that may be used to build, corroborate, integrate or enhance an argument. The conscious choice of limiting the analysis to carefully selected keywords, rather than extended thematic dictionaries or more advanced techniques, while certainly oversimplifying, provides meaningful results that can be understood and interpreted by any informed reader. As the number of cases increases, and the questions asked become less linear, introducing more formal or complex methods of analysis becomes a necessity.

Limitations

The approach outlined in this article may be useful in a wide set of circumstances. However, it has substantial limitations that should be fully considered. The availability of content from relevant sources and for relevant periods of time is a fundamental restriction. However, the most important limitation is substantive: researchers should not demand of basic word frequency analysis more than it can really say.

Availability of content

Analysis of word frequency based on time series is at its most useful when lengthy periods of time are included. For example, a dataset should ideally include a substantial amount of data related to the periods both before and after key events.

Without a doubt, an ever increasing amount of textual content is available on-line. However, the relative novelty of the internet leads to the fact that very rarely are materials published before 2000 readily available. Moreover, it has been common for websites to discard older contents when moving the website to a new platform or content management system. As a consequence, the currently available version of a website often includes only content published starting with 2010 or later (see also Table 1). This is a substantial limitation, even for research focused on contemporary events. For example, it is difficult to find relevant local sources with content published both before and after a relatively recent event such as the August 2008 war in South Ossetia.

Institutional websites tend to have longer life-spans. However, sometimes political rather than technical considerations limit the availability of older contents. For example, when a new president has been elected in countries such as Georgia (in 2013) and Ukraine (in 2014), all the contents published under the previous president have been deleted, thus highlighting a moment of discontinuity with the previous leadership of the country.[17]

17 Or rather, it may be argued, shedding light on what seems to be a fundamental misunderstanding about the continuity of democratic institutions.

Retrieval of desired contents and metadata

There are multiple possible approaches to extracting textual content and metadata from a website for tech-savvy researchers. However, this may be a time-consuming process that requires ad hoc solutions and dedicated scripts for each website. All data presented in this article have been collected using Castarter – Content Analysis Starter Toolkit for R, a package originally developed by this author for the open source R software. Castarter gives the possibility to users with no programming skills to extract textual content and metadata from a given section of a website, thanks to the fact that most modern content management systems distribute content using one of a few standard approaches.

To further test the feasibility of creating datasets from multiple websites, in October 2015 this author has retrieved with Castarter all press-releases of the websites of the presidents of the 15 former Soviet republics, and none of them posed insurmountable challenges. However, things may not go so smoothly with non-institutional websites, and in particular extracting content from older websites may be more complicated. For example, the website of Transnistria's state news agency offers free access to all of its content in a format that can easily be extracted by Castarter for all news published since 2012. Previous content all the way back to 1999 is available, but in a less accessible format. They have been successfully extracted by this author, but this required developing a customised parsing solution.

Textual content can be usefully analysed as described in this paper only if it is currently freely available on the internet and can be extracted in a reasonable amount of time. Extracting content for recent years from modern, structured websites (such as most institutional websites or blogs) is usually a relatively straightforward process. However, if relevant data is not currently available online, or if it is overtly complex to extract, the approach outlined in this article becomes infeasible.

Substantive issues and assumptions

Finally, the most important limitations regard substantive issues. The approach outlined in this article relies heavily on word frequency, which is the most basic tool of quantitative content analysis. The rudimentary nature of the process of counting the number of occurrences of a given term requires that the researcher is fully aware of the context in which that term has been used, the existence of alternative expressions to refer to the same concept, and the polysemantic nature of some words.
Accordingly, as has been highlighted, this approach may be useful in the preliminary phase of the research and as a tool to provide additional evidence or information on specific issues. However, the researcher should use such data with caution, without overestimating their explanatory power, and mostly in combination with other methods.

Conclusions

Transforming the textual content of a website into a structured dataset can be a useful preliminary step for informing qualitative research. On the one hand, it allows to search, subset and browse through the content of a website in an effective, and, if useful, in an exhaustive way. Such an approach may be used to give structure and formalise a part of the research process that is now commonplace, even if rarely explicitly described: looking for information online through search engines or by browsing relevant websites. It is argued here that whenever possible unstructured online searches should be substituted by (or integrated with) an organic analysis that can be formally described among the steps of the research process.

Given the relative simplicity of creating new datasets and conducting basic word frequency analysis, this approach may be used to provide additional evidence in support of an argument, or to enhance it, even if this may eventually become simply a footnote in a paper. Such a footnote may strengthen the argument being made and provide substantial additional evidence. For example,

a statement such as 'the crisis in Ukraine has dominated the news on Russia's state television in 2014 and 2015', could be accompanied by the following footnote: 'An analysis of all news items published on the website of Russia's state owned First Channel since the beginning of Vladimir Putin's third mandate as president (N = 111,917, time frame: 7 May 2012 – 31 December 2016) shows that more than one quarter of news stories made reference to Ukraine for each single month between March 2014 and April 2015, with almost half of news items mentioning Ukraine at least once in the months of March 2014, July 2014 and February 2015.' Additional information on how such data has been obtained or relevant graphs (see for example Illustration 15) could be provided in an appendix or a dedicated repository.

Fundamentally, such analyses should not be restricted to pre-existing datasets, but may be based on new datasets created ad hoc from a relevant online source. Scholars, for example, have started to include references to Google Ngram – a tool that allows to explore word frequency of a term in printed sources – to illustrate their arguments about the frequency of references to individuals or concepts (Elgie, 2015; Etkind, 2013). The same can be done whenever relevant data are currently available on a single website, or a clearly defined set of websites, which is not uncommon in area studies. For example, in their article on Georgia's European identity, Ó Beacháin & Coene (2014) could have used this approach to offer additional evidence on the prominence of specific attributes of 'Europeanness' in Saakashvili's public rhetoric, and Loda (2016) could have used it to enrich her analysis of key aspects of Azerbaijan's public diplomacy as they emerge in presidential speeches published on the official website of Azerbaijan's president. Use-case scenarios, however, are not limited to official websites or established media outlets: analysing datasets created from selected blogs or forums may be an inobstrusive and effective method for finding mundane and otherwise difficult to notice forms of self-expression that can be further investigated, for example, with an ethnographic approach (Seliverstova & Pawlusz, 2016).

More in general, analyses of word frequency may become a constituent part of the preliminary phase of qualitative research in a number of circumstances; it may be employed to choose cases or key moments, and to strengthen claims that such choices were not the results of the researcher's bias, as well as to provide additional evidence in preparation of interviews or further research with other methods. It is a non-obtrusive and replicable data collection method that may be usefully introduced in the toolbox of many researchers. Finally, once the limitations of the method are fully considered, this approach can be used to explore comparisons among cases, 'as one tool among many in the comparativist toolkit' (Kevlihan, 2013), including in circumstances where no standard measure for comparison exists.

The increased availability of large amounts of structured textual content freely available on-line and of software packages that allow for their analysis has drawn increasing attention towards quantitative content analysis. Often, studies based on quantitative content analysis make use of complex models that are overtly difficult to understand for the uninitiated. This leads to a situation in which scholars with an interest in the substantive issues analysed are not able to judge independently the reliability and the actual meaning of the results presented. At the same time, scholars that are less tech-savvy or not accustomed to use quantitative methods, tend to overlook altogether the vast amount of structured contents that has become available on-line in recent years, or to explore it serendipitously.

This needs not be the case. 'Noting trends and changes in content' (Albig 1938, 349) may still be considered a key element of content analysis; not every scholar using quantitative content analysis should necessarily strive to use more advanced techniques or complex machine learning models. Scholars who are not familiar with advanced content analysis techniques may intuitively understand analyses based on word frequency and appreciate their inherent limitations. Finally, the availability of software packages such as Castarter substantially lowers the threshold of technical competence required for creating new datasets, exploring them, and sharing the results.[18] As the tools to conduct such analysis become easier to use, researchers may well consider including them routinely in their repertoire.

18 New datasets, or the scripts required to create them, can easily be shared easily on-line, further facilitating the process for less tech-savvy researchers.

References

Albig, W. (1938). The Content of Radio Programs, 1925-1935. *Social Forces*, *16*(3), 338–349. https://doi.org/10.2307/2570805

Apsnypress.info, (21 February 2012). "Aleksandr Ankvab ne pitaet illyuzii po povody uluchsheniya otnoshenii s Evrosoyuzom", , http://www.apsnypress.info/news/aleksandr-ankvab-ne-pitaet-illyuziy-po-povodu-uluchsheniya-otnosheniy-s-evrosoyuzom-/

Apsnypress.info, (26 June 2012). "Aleksandr Ankvab: tem, kto khochet sotrudnichat' s Abkhaziei, sleduet brat' primer s Rossii", , http://www.apsnypress.info/news/aleksandr-ankvab-tem-kto-khochet-sotrudnichat-s-abkhaziey-sleduet-brat-primer-s-rossii-/;

Apsnypress.info, (11 September 2012). "Aleksandr Ankvab prinyal posla Niderlandov v Armenii i Gruzii", , http://www.apsnypress.info/news/aleksandr-ankvab-prinyal-posla-niderlandov-v-armenii-i-gruzii-/;

Baker, P., & McEnery, T. (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.

Benoit, K., & Nulty, P. (2016). quanteda: Quantitative Analysis of Textual Data (Version 0.9.8). Retrieved from https://CRAN.R-project.org/package=quanteda

Bouchet-Valat, M., & Bastin, G. (2013). RcmdrPlugin.temis, a Graphical Integrated Text Mining Solution in R. *The R Journal*, *5*(1), 188–196.

Comai, G. (2015a). Post-Soviet de facto states online. In W. Schreiber & M. Kosienkowski (Eds.), *Digital Eastern Europe*. Wrocław: KEW.

Comai, G. (2015b, November 3). Word frequency of 'Ukraine', 'Crimea', and 'Syria' on Russia's First Channel. Retrieved from http://www.giorgiocomai.eu/2015/11/03/word-frequency-of-ukraine-crimea-and-syria-on-russias-first-channel/

Elgie, R. (2015, September 7). Maurice Duverger, semi-presidentialism, and explaining. Retrieved from http://presidential-power.com/?p=3737

Etkind, A. (2013). Mourning and melancholia in Putin's Russia. In E. Rutten, J. Fedor, & V. Zvereva (Eds.), *Memory, Conflict and New Media: Web Wars in Post-Socialist States*. Routledge.

Feinerer, I., Hornik, K., & Meyer, D. (2008). Text mining infrastructure in R. *Journal of Statistical Software*, *25*(5), 1–54.

Fetters, M. D., & Freshwater, D. (2015). Publishing a Methodological Mixed Methods Research Article. *Journal of Mixed Methods Research*, *9*(3), 203–213. https://doi.org/10.1177/1558689815594687

Franzosi, R. (Ed.). (2008). *Content analysis*. London: SAGE.

Fürsich, E., & Robins, M. (2002). Africa.com: the self-representation of sub-Saharan nations on the World Wide Web. *Critical Studies in Media Communication*, *19*(2), 190–211. https://doi.org/10.1080/07393180216557

Ker-Lindsay, J. (2015). Engagement without recognition: the limits of diplomatic interaction with contested states. *International Affairs*, *91*(2), 267–285. https://doi.org/10.1111/1468-2346.12234

Kevlihan, R. (2013). Designing Social Inquiry in Central Asia – A Case Study of Kyrgyzstan and Tajikistan. *Studies of Transition States and Societies*, *5*(1).

King, C. (2001). The Benefits of Ethnic War: Understanding Eurasia's Unrecognized States. *World Politics*, *53*(4), 524–552. https://doi.org/10.2307/25054164

Krippendorff, K. (2004). *Content analysis: an introduction to its methodology*. Thousand Oaks, Calif.: Sage.

Kutter, A., & Kantner, C. (2012). Corpus-Based Content Analysis: A Method for Investigating News Coverage on War and Intervention. International Relations Online Working Paper, 2012(01), February 2012, Stuttgart: Stuttgart University. Retrieved from http://www.uni-stuttgart.de/soz/ib/forschung/IRWorkingPapers/IROWP_Series_2012_1_Kutter_Kantner_Corpus-Based_Content_Analysis.pdf

Laboratorul Pentru Analiza Conflictului Transnistran. (2013). Resurse Transnistrene Online. Retrieved August 12, 2015, from http://lact.ro/index.php/proiecte/81-resurse-transnistrene-online

Law of Georgia on Occupied Territories, Law n. 431 (2008). Retrieved from https://matsne.gov.ge/en/document/view/19132

Loda, C. (2016). Azerbaijan, Foreign Policy and Public Diplomacy. *Irish Studies in International Affairs*, *27*, 39–55. https://doi.org/10.3318/isia.2016.27.7

McMillan, S. J. (2000). The microscope and the moving target: The challenge of applying content analysis to the World Wide Web. *Journalism & Mass Communication Quarterly*, *77*(1), 80–98.

Mohammed, S. N. (2004). Self-Presentation of Small Developing Countries on the World Wide Web: A Study of Official Websites. *New Media & Society*, *6*(4), 469–486. https://doi.org/10.1177/146144804044330

Moldova's MFA. 2013. "Ministry of Foreign Affairs and European Integration of the RM." *Moldova's MFA/Archive.Org*. July 22. http://web.archive.org/web/20130722140536/http://www.mfa.gov.md/start-page-en/.

Moran-Ellis, J., Alexander, V. D., Cronin, A., Dickinson, M., Fielding, J., Sleney, J., & Thomas, H. (2006). Triangulation and integration: processes, claims and implications. *Qualitative Research*, *6*(1), 45–59. https://doi.org/10.1177/1468794106058870

Ó Beacháin, D., & Coene, F. (2014). Go West: Georgia's European identity and its role in domestic politics and foreign policy objectives. *Nationalities Papers*, *42*(6), 923–941. https://doi.org/10.1080/00905992.2014.953466

Ó Beacháin, D., Comai, G., & Tsurtsumia-Zurabashvili, A. (2016). The secret lives of unrecognised states: Internal dynamics, external relations, and counter-recognition strategies. *Small Wars & Insurgencies*, *27*(3), 440–466. https://doi.org/10.1080/09592318.2016.1151654

O'Loughlin, J., Kolossov, V., & Toal, G. (2011). Inside Abkhazia: Survey of Attitudes in a De Facto State. *Post-Soviet Affairs*, *27*(1), 1–36. https://doi.org/10.2747/1060-586X.27.1.1

R Core Team. (2016). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from http://www.R-project.org/

Riffe, D., Lacy, S., & Fico, F. (2005). Analyzing media messages using quantitative content analysis in research. Mahwah, N.J.: Lawrence Erlbaum.

Schreier, M. (2012). *Qualitative Content Analysis in Practice*. London: SAGE.

Seliverstova, O., & Pawlusz, E. (2016). Everyday Nation-Building In The Post-Soviet Space. Methodological Reflections. *Studies of Transition States and Societies*, *8*(1).

Silge, J., & Robinson, D. (2016). tidytext: Text Mining and Analysis Using Tidy Data Principles in R. *The Journal of Open Source Software*, *1*(3). https://doi.org/10.21105/joss.00037

Wickham, H. (2009). *Ggplot2: elegant graphics for data analysis*. New York: Springer.

Acknowledgement

This research was supported by a FP7/Marie Curie ITN action. Grant agreement N°: 316825

Appendix 1

Name of website	Earliest publication available	Total number of publications	Average number of publications per day
Abkhazia MFA	2012-04-16	742	0.48
Armenia MFA	2007-01-09	1,874	0.54
Azerbaijan MFA	2012-03-05	1,050	0.67
Georgia MFA	2015-01-05	1,057	1.95
Iceland MFA	1995-05-15	821	0.11
Iran MFA	2001-02-20	10,182	1.82
Kosovo MFA	2008-12-03	1,332	0.27
Malta MFA	2013-03-15	412	0.34
Moldova MFA	2007-03-07	965	0.28
Montenegro MFA	2013-07-25	497	0.64
Nagorno Karabakh MFA	2008-11-05	486	0.17
North Cyprus MFA	2013-02-12	361	0.29
Russia MFA	2003-01-04	16,236	3.30
San Marino MFA	2007-04-20	267	0.08
South Ossetia MFA	2010-04-19	347	0.15
Transnistria MFA	2009-01-06	1,335	0.49
Ukraine MFA	2010-01-12	1,813	0.77

Table 1: List of websites included in the analysis and summary statistics updated to 30 June 2016.

Appendix 2

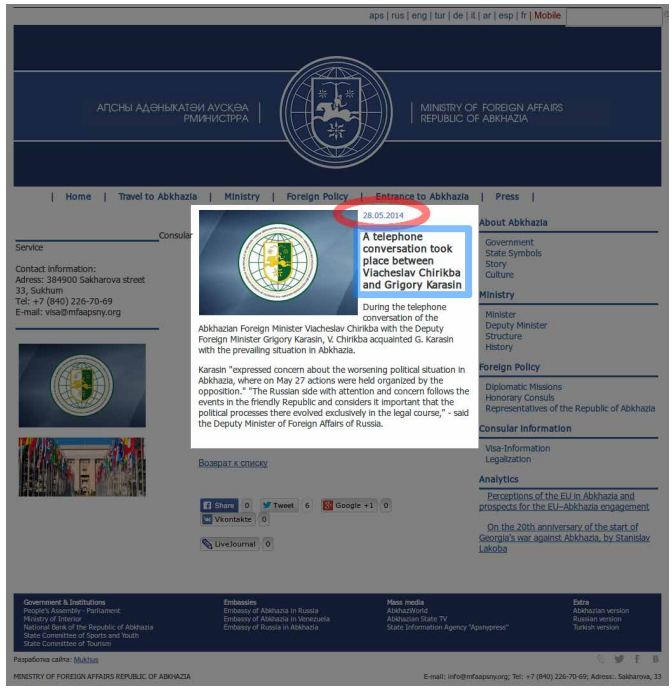

Illustration 1: Extracting textual contents and metadata from a web page

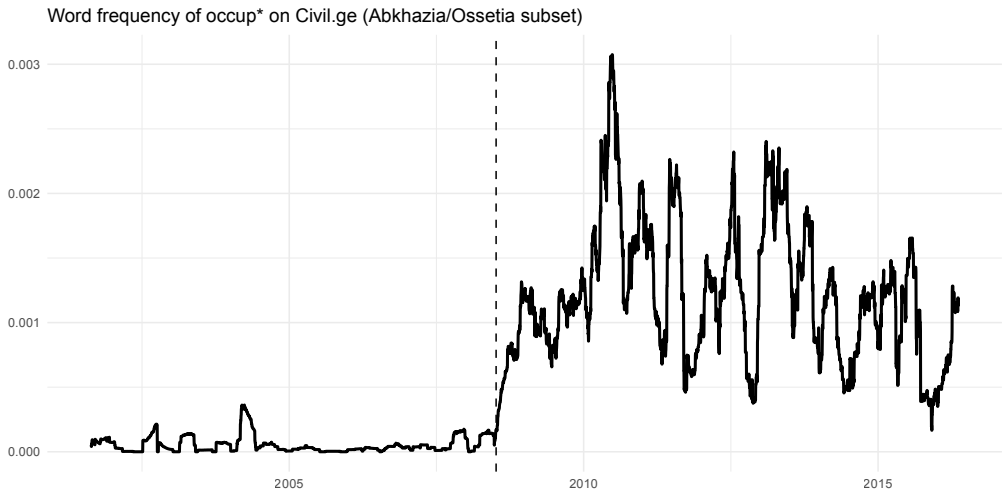

Illustration 2: Word frequency of 'occupation' ('occup*') in the articles published on the English language version of Civil.ge before the end of June 2016 that mentioned either 'Abkhazia' or 'Ossetia' (N = 8,671, out of a total of 27,447 articles). The frequency is calculated on a rolling average of 90 days to enhance readability. The vertical dotted line highlights the first day of the war in South Ossetia on 7 August 2008.

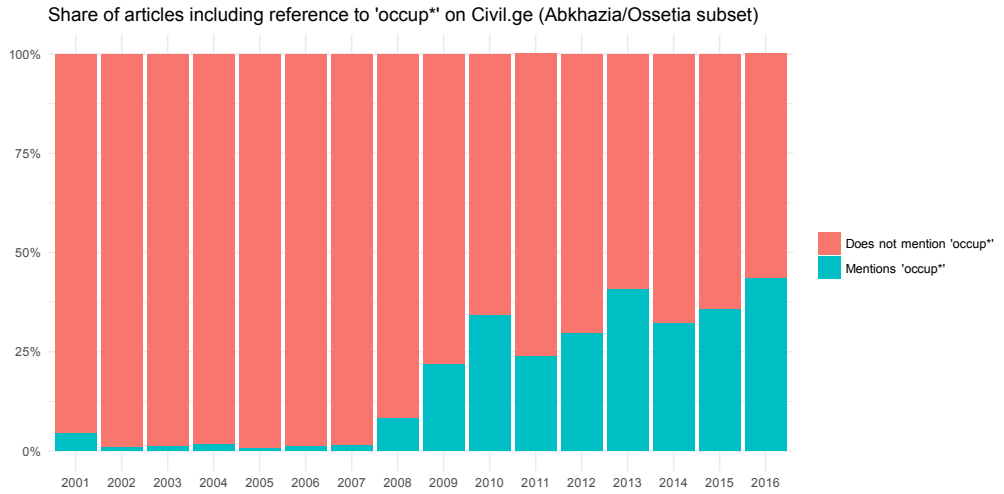

Illustration 3: Share of articles including reference to 'occupation' ('occup*') in the articles published on the English language version of Civil.ge before the end of June 2016 that mention either 'Abkhazia' or 'Ossetia' (N = 8,671, out of a total of 27,447 articles).

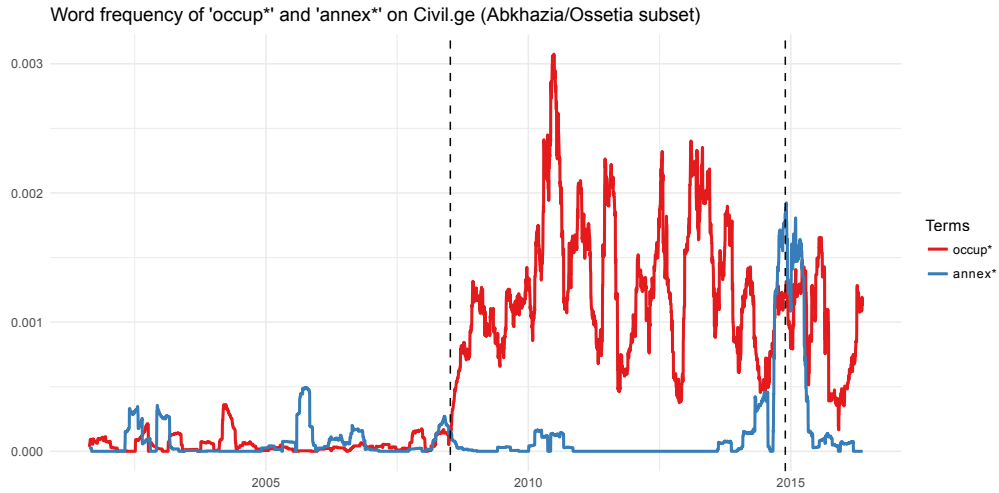

Illustration 4: Word frequency of 'annexation' and 'occupation' in the articles published on the English language version of Civil.ge before the end of June 2016 that mentioned either 'Abkhazia' or 'Ossetia' (N = 8,671, out of a total of 27,447 articles). The frequency is calculated on a rolling average of 90 days to enhance readability. The first vertical dotted line highlights the first day of the war in South Ossetia on 7 August 2008, the second vertical line highlights the signature of the 'strategic partnership' treaty between Russia and Abkhazia on 24 November 2014.

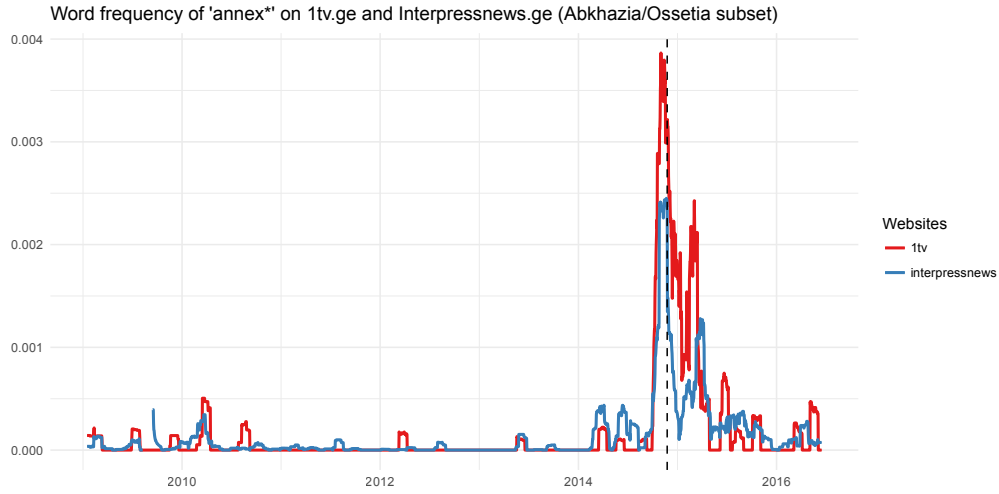

Illustration 5: Word frequency of 'annexation' and 'occupation' in the articles published on the Georgian-language version of 1tv.ge (Georgia's state owned First Channel) and Interpressnews.ge (Georgia's largest news website) between January 2009 and June 2016 that mentioned either 'Abkhazia' , 'Ossetia' or 'Tskhinvali' (N = 22,172 out of a total of 382,603 articles, or 4,894 out of 116,014 from 1tv.ge and 17,278 out of 266,589 from Interpressnews.ge). The frequency is calculated on a rolling average of 90 days to enhance readability. The vertical dotted line highlights the signature of the 'strategic partnership' treaty between Russia and Abkhazia on 24 November 2014.

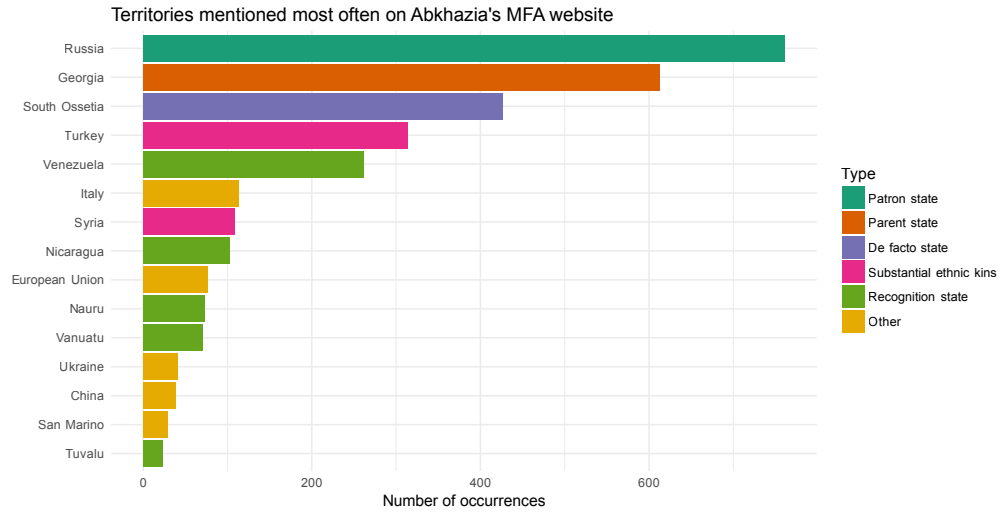

Illustration 6: Countries mentioned most often on Abkhazia's MFA website, divided by type of country. Results based on number of references to any UN member state (plus de facto states, and European Union for reference) in all press releases available on Abkhazia's MFA current website available starting with April 2012 and until 30 June 2016 (N = 742).

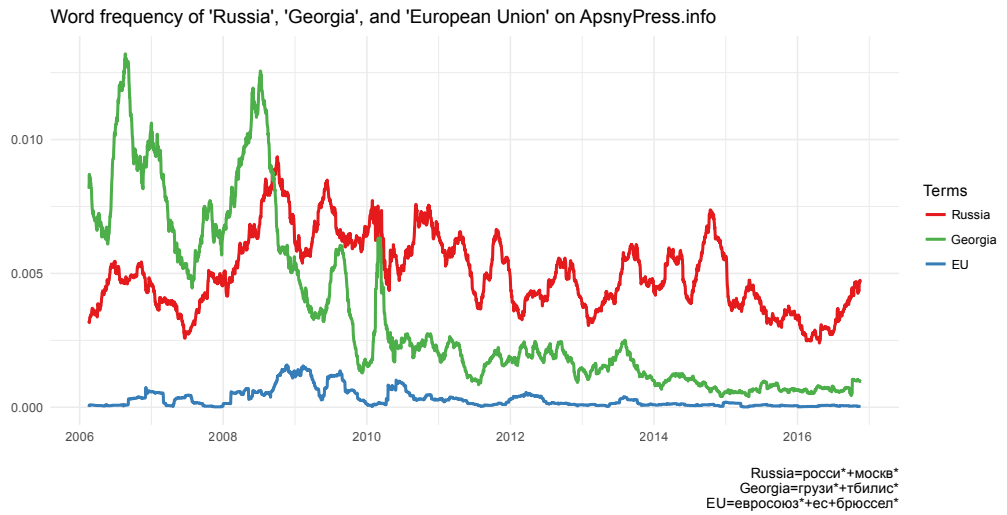

Illustration 7: Word frequency of 'Russia, 'Georgia', and 'European Union' in all articles published on the website of Abkhazia's news agency ApsnyPress between 1 January 2006 and 31 December 2016 (N=25,618), calculated on a rolling average of 90 days for clarity.

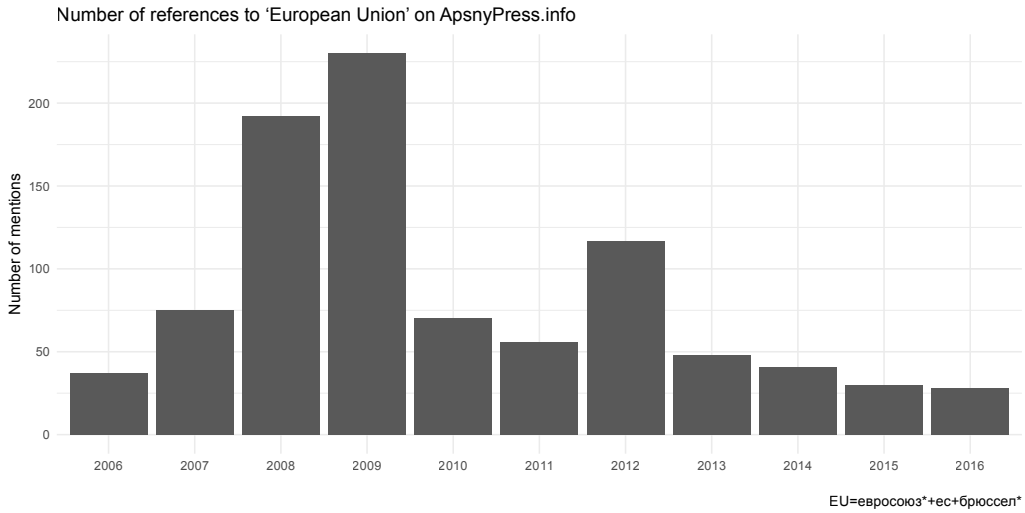

Illustration 8: Number of references to 'European Union' in all articles published on the website of Abkhazia's news agency ApsnyPress between 1 January 2006 and 31 December 2016 (N=25,618), subdivided by year for clarity.

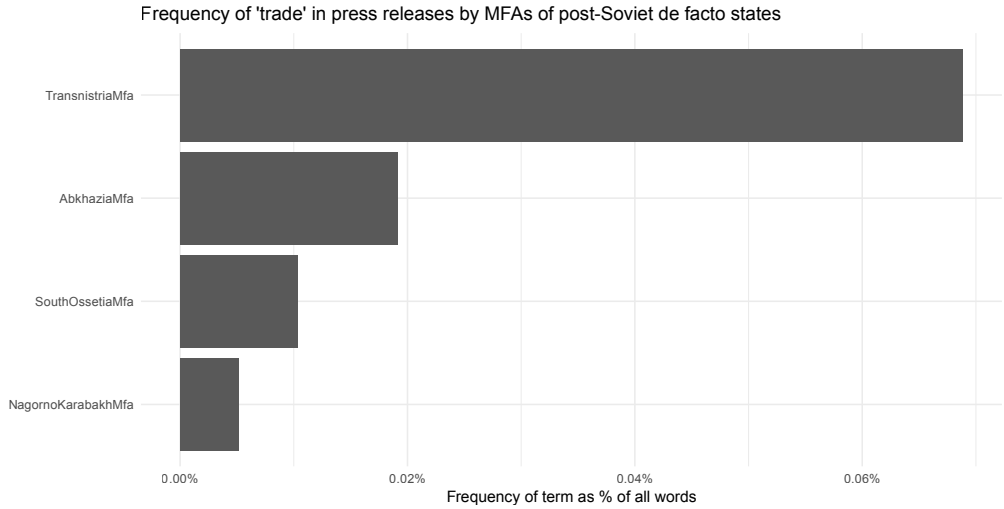

Illustration 9: Word frequency of 'trade' in the items published on the websites of the MFAs of post-Soviet de facto states between 1 July 2012 and 30 June 2016 (N = 2,234)

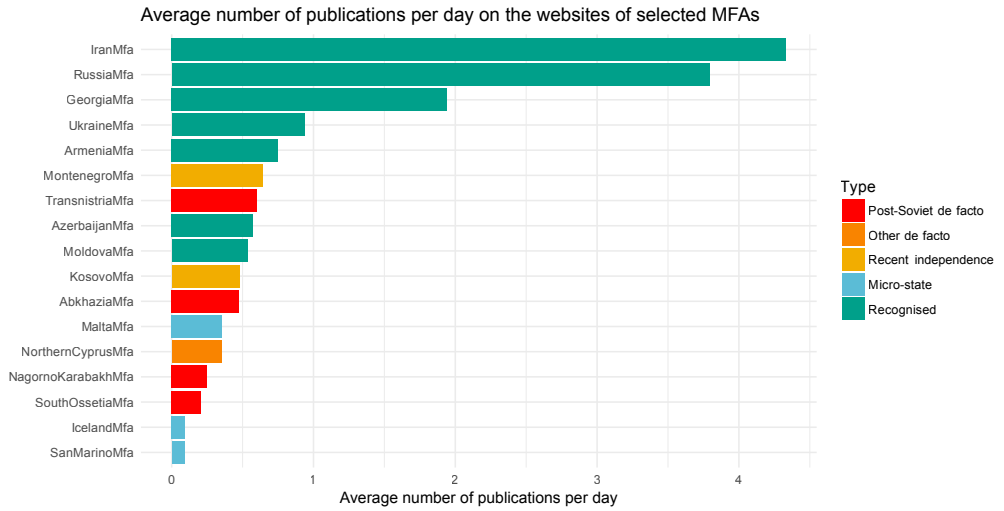

Illustration 10: Average number of publications per day on the websites of selected MFAs between 1 July 2013 and 30 June 2016 (N = 16,584)

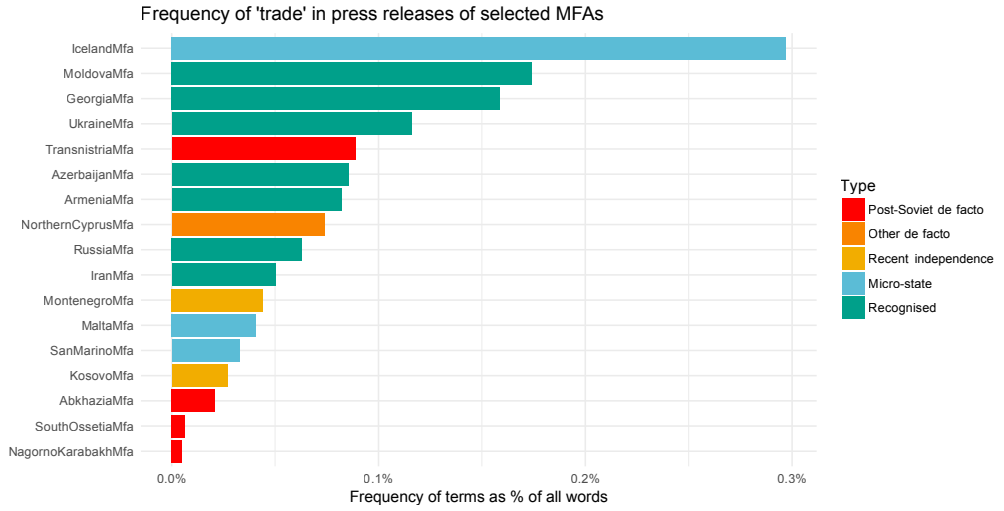

Illustration 11: Word frequency of 'trade' in the items published on the websites of selected MFAs between 1 July 2013 and 30 June 2016 (N = 16,584)

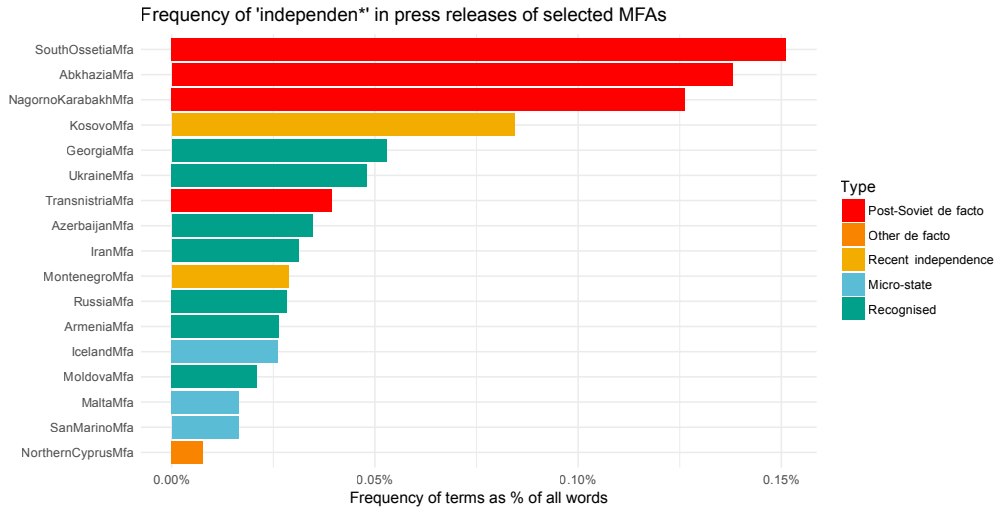

Illustration 12: Word frequency of 'independence' ('independen*') in the items published on the websites of selected MFAs between 1 July 2013 and 30 June 2016 (N = 16,584)

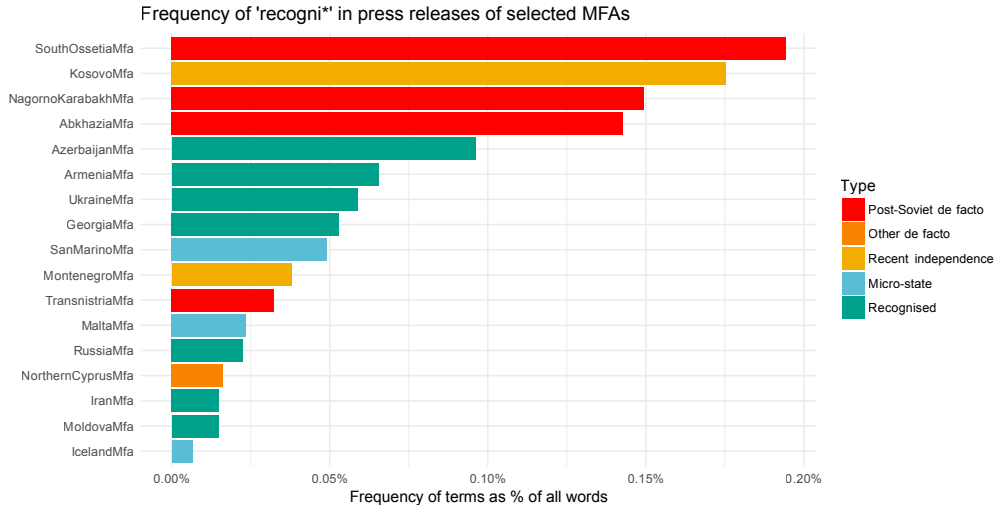

Illustration 13: Word frequency of 'recognition' ('recogni*') in the items published on the websites of selected MFAs between 1 July 2013 and 30 June 2016 (N = 16,584)

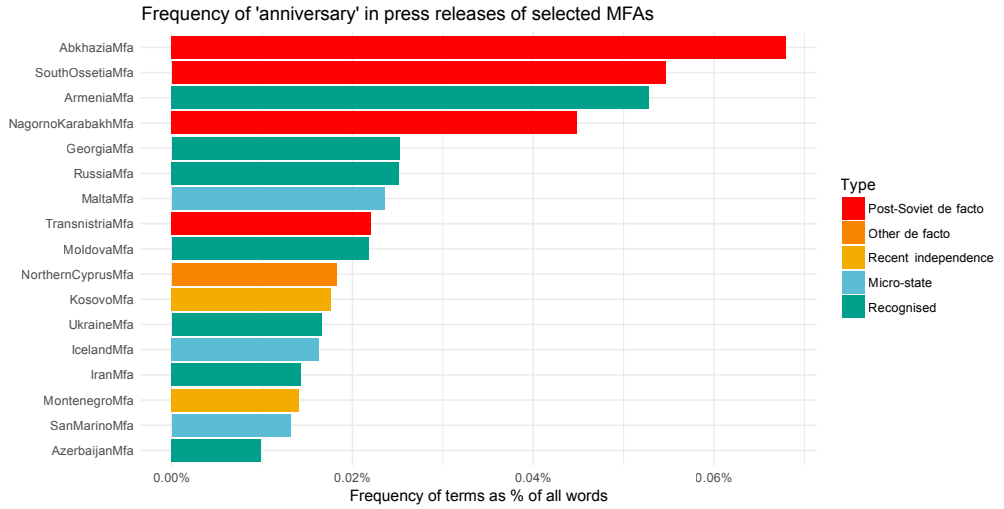

Illustration 14: Word frequency of 'anniversary' in the items published on the websites of selected MFAs between 1 July 2013 and 30 June 2016 (N = 16,584). It is worth highlighting that in the period under analysis Armenia commemorated the centenary of the Armenian genocide.

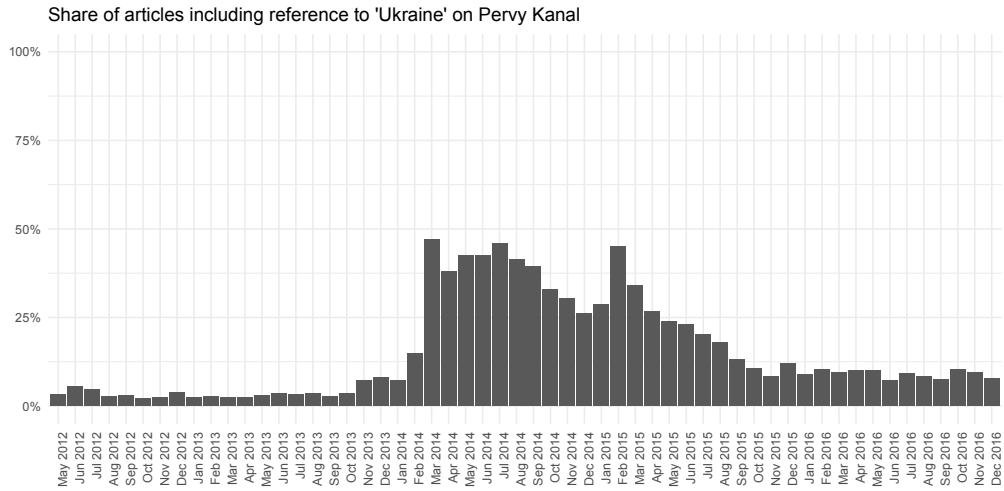

Illustration 15: Share of news items published on the website of Russia's First Channel - 1tv.ru - including reference to 'Ukrain*' since the beginning of Vladimir Putin's third mandate as president (N = 111,917, time frame: 7 May 2012 – 31 December 2016). For more details see Comai (2015b).