

# The GriCo Project: A Web Corpus of the Italian Populist Movimento 5 Stelle

Matteo Di Cristofaro, Dario Del Fante, Federica Formato,  
Elena Valvason, Virginia Zorzi, Angela Zottola

# Overview

- Our team
- The political scenario
- An introduction to the GriCo corpus
- Three exploratory pilot studies:
  - Diachronic variation in complexity and lexical diversity
  - Diachronic variation in the use of the modal verb form *dobbiamo*
  - Keywords
- Future developments



# Research team



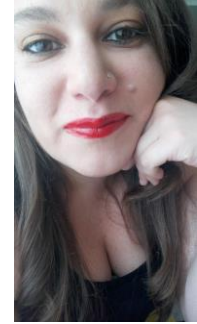
Dario Del Fante

University of Padova/ Sussex University



Matteo Di Cristofaro

University of Modena and Reggio  
Emilia



Federica Formato

University of Brighton



Elena Valvason

University of Pavia/University of Bergamo



Virginia Zorzi

University of Padova



Angela Zottola

Lancaster University

# Movimento 5 Stelle: a novel case in the populist European scenario

- Gained **25%** in the 2013 elections and **32%** in 2018 election
- M5S speaks on behalf of '**the people**' and claims to embody the popular will
- Beppe **Grillo**: founder and former leader
- Luigi **Di Maio**: current leader (Deputy PM and Minister of Minister of Economic Development, Labour and Social Policies)



# Movimento 5 Stelle: a novel case in the populist European scenario

- The web is a *super medium*: it allows for direct democracy (power to the people) and it's a mythical panacea (Natale & Ballatore 2014: 113)
- The blog **is** their manifesto and “fulfills the function of place for debate, party's [sic] house organ, and voting platform” (Maccaferri 2018: 100)
- Previous studies on the language used by M5S
  - epithets used by Grillo on the blog between 2008 and 2015 to refer to Berlusconi and three successive centre-left leaders (Bortoluzzi & Semino 2016)



# The GriCo corpus – Why?

- A scarcity of Italian corpora on political language
- Increasing relevance of populism in the European scenario (Wodak & Krzyżanowski 2017)
- Freely available large amount of data

**FACILITATORI**  
SCELTI TRA I PRIMI 10 VOTATI DAGLI ISCRITTI

**FACILITATORI**  
SCELTI TRA I PRIMI 15 VOTATI DAGLI ISCRITTI

**8 FACILITATORI**  
SCELTI TRA I PRIMI 20 VOTATI DAGLI ISCRITTI



19 luglio, 2019

MoVimento 5 Stelle

## La proposta di organizzazione regionale del MoVimento 5 Stelle



**LUIGI DI MAIO**

CAPO POLITICO DEL MOVIMENTO 5 STELLE

21 COMMENTI



122



23

Nell'ultimo video abbiamo parlato del Team del Futuro, dei 12 facilitatori nazionali più i 6 ruoli di organizzazione nazionale. In questo secondo v...

- Named as *beppegrillo.it*, it changed into *ilblogdellestelle.it* in 2018
- The dataset contains posts written between January 2005 and December 2018 users' comments and replies written between January 2005 and April 2019

# Collection procedure

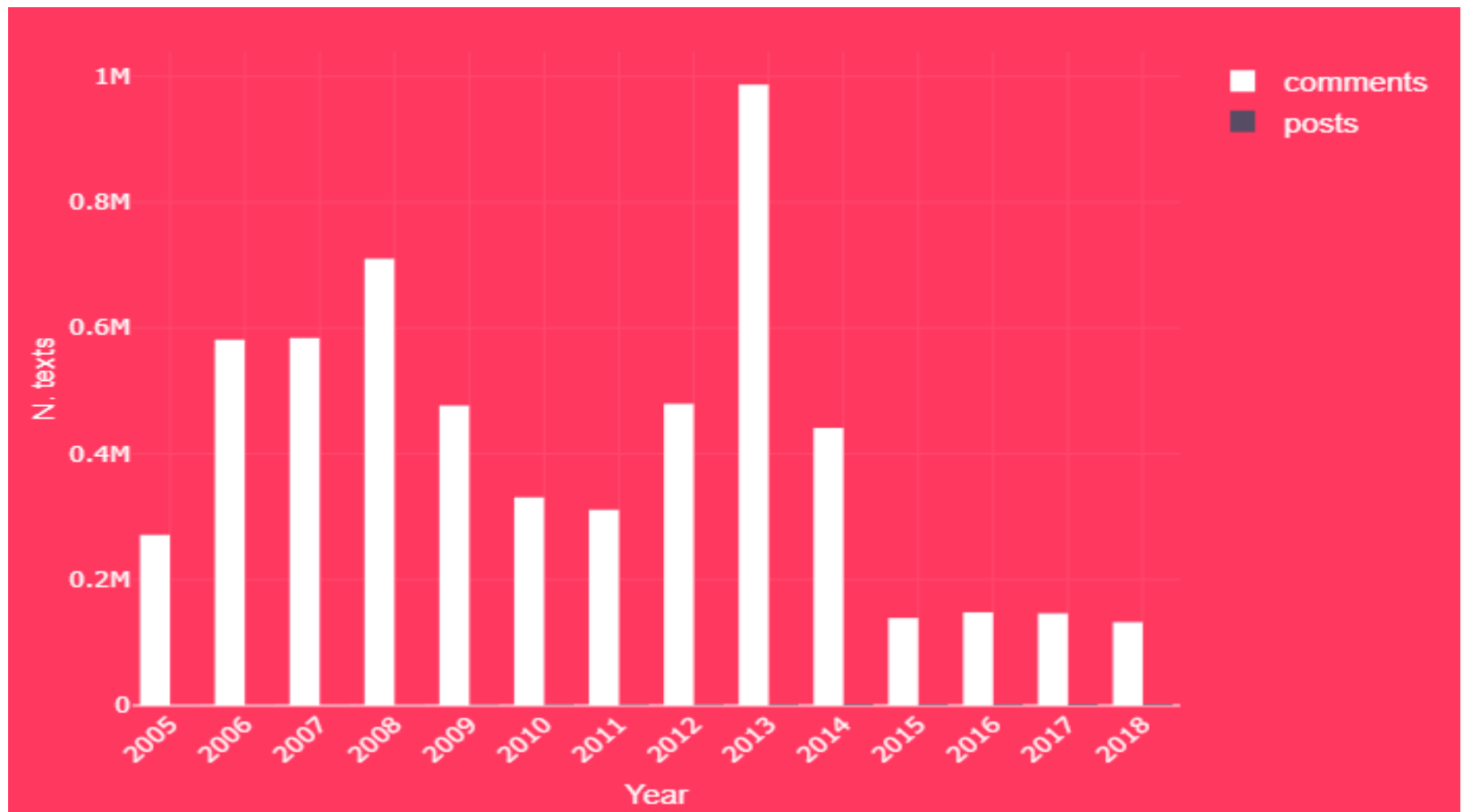
- The raw *html* data was collected with custom Python scripts (using *requests*, *BeautifulSoup* and *lxml*)
- Data processing was performed using custom Python scripts to extract the relevant pieces of information (text and metadata) and output them in different formats (plain text and XML)
- POS tagging was added to the XML files; two versions are available, using two different taggers (TreeTagger and spacy.io)



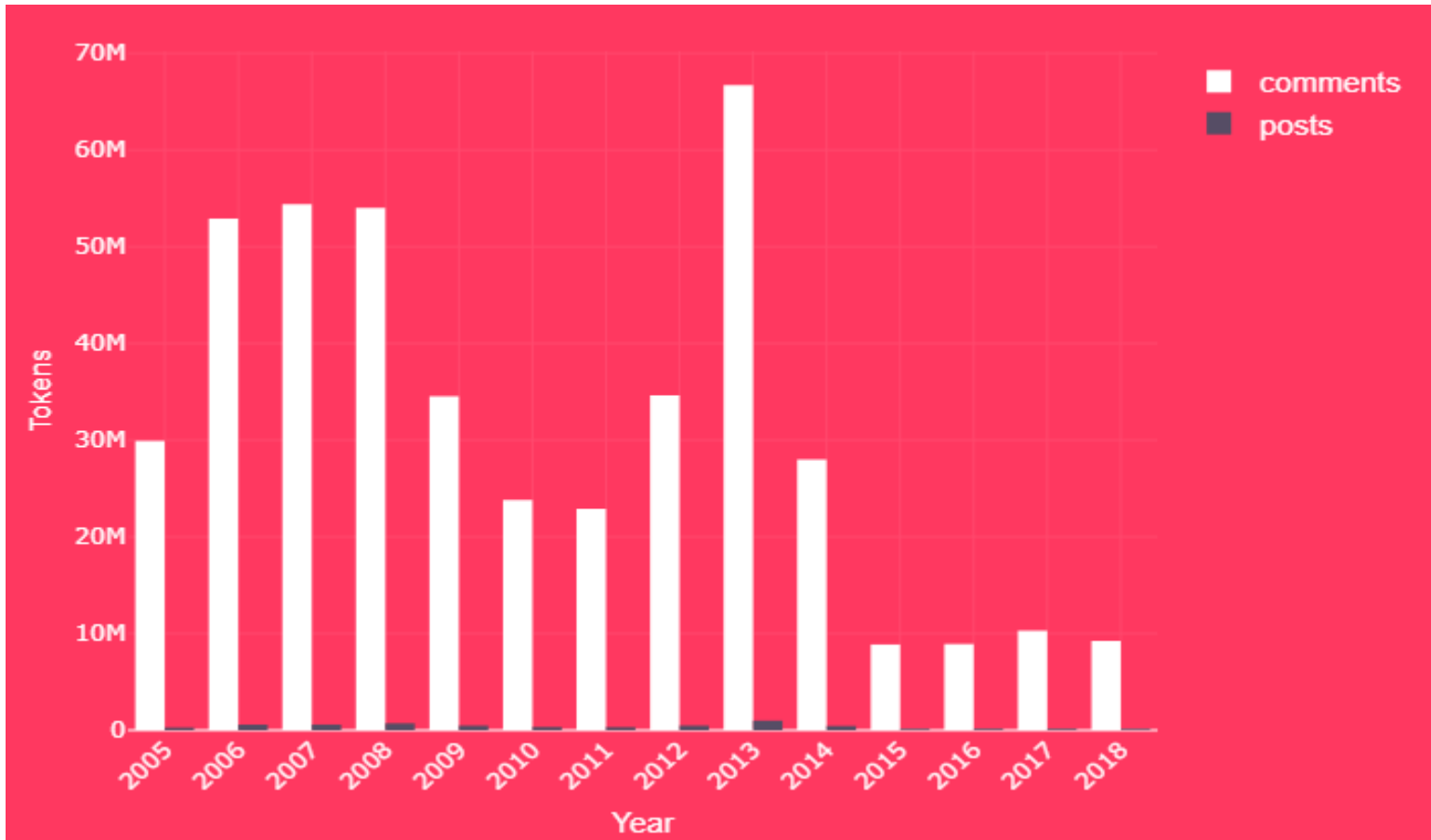
# GriCo: an overview

<b>(sub)corpus</b>	<b>n. texts</b>	<b>tokens</b>
Posts	15,437	7,324,806
Comments + Discussions	5,739,758	439,173,428
GriCo	5,755,195	446,498,234

<https://grico.gitlab.io/>



Number of texts, divided by year and category (posts or comments)



Number of tokens, divided by year and category (posts or comments)

# Metadata: an example

The XML corpus preserves the website's structure and part of its metadata. Each page is saved to a separate XML files, containing the posts' text and its comments. The metadata is structured as follows:

```
<?xml version='1.0' encoding='utf-8' standalone='yes'?>

<text collected="2019-04-23T17:10:36" id="15158" title="TITLE" url="FULL_URL">

  <u author="NAME" blogid="15158" childof="none" date_d="DD" date_m="MM" date_y="YYYY" tags="TAGS"
type="post"></u>

  <u author="NAME" blogid="82764" childof="15158" date_d="DD" date_m="MM" date_y="YYYY" tags="none"
type="comment"> </u>

  <u author="NAME" blogid="82889" childof="82764" date_d="DD" date_m="MM" date_y="YYYY" tags="none"
type="discussion"> </u>

</text>
```

# GriCo Sampler

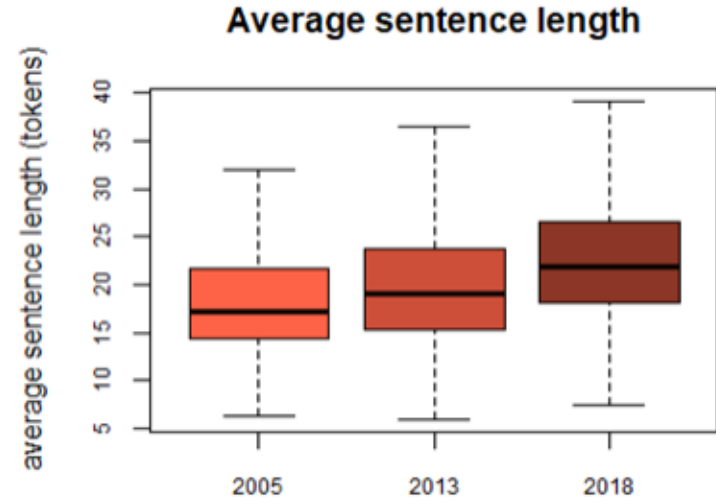
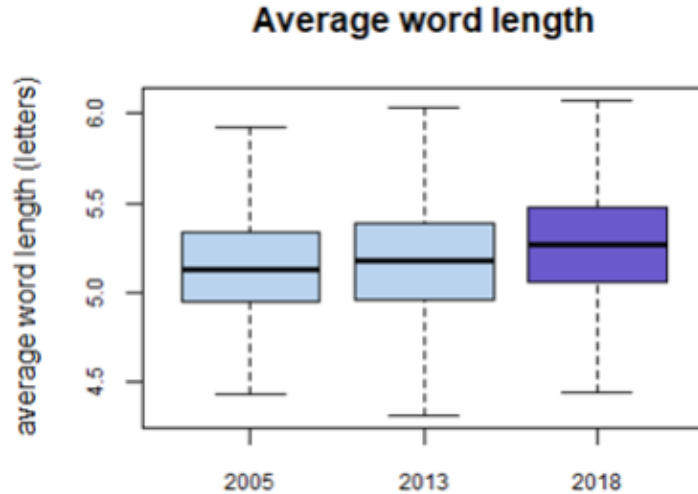
3 sampling points in time (McEnery & Baker 2016):

- **2005 sample:** the first year the blog was active
- **2013 sample:** when they were firstly elected as part of the Italian Parliament
- **2018 sample:** when they were the most voted party in the Italian General Elections

# Pilot study 1: complexity and lexical diversity in a diachronic perspective (1)

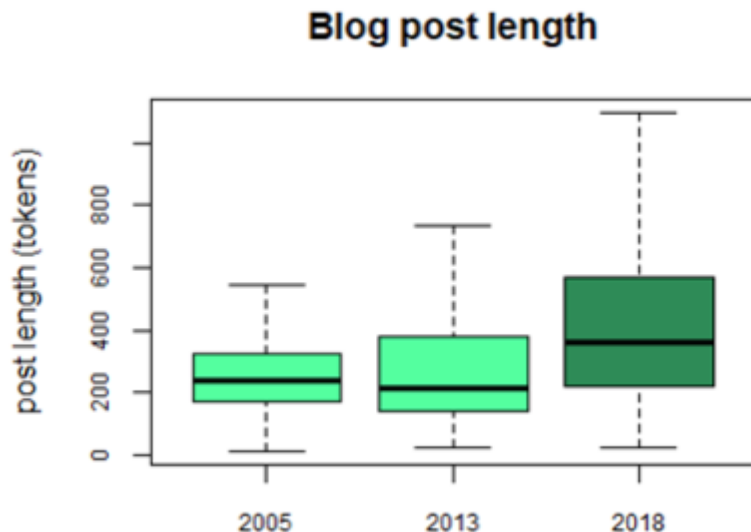
	2005	2013	2018	
Complexity	Avg. sentence length	18.54	21.25	23.25
	Sent. length st. dev.	6.59	11.93	8.51
	Avg. word length	5.16	5.18	5.28
	Word length st. dev.	0.35	0.33	0.32
Text length	Avg. post length	274.57	361.26	463.33
	Post length st. dev.	177.59	492.72	404.93
Lexical diversity	Avg. MATTR	0.77	0.77	0.78
	MATTR st. dev.	0.04	0.04	0.04

# Pilot study 1: Complexity and lexical diversity in a diachronic perspective (2)



- Change in colour → statistically significant difference ( $p < 0.05$  resulting from the Wilcoxon rank sum test)
- Words become significantly longer from 2005 - 2013 to 2018
- Sentences become significantly longer from 2005 to 2013 and from 2013 to 2018

## Pilot study 1: Complexity and lexical diversity in a diachronic perspective (3)



- Change in colour → statistically significant difference ( $p < 0.05$  resulting from the Wilcoxon rank sum test)
- Blog posts become significantly longer from 2005-2013 to 2018



## Pilot study 2: The modal verb form *dobbiamo* in diachrony (1)

- *dobbiamo* (present tense, indicative mood) ‘we must/have to/ought to’ generally functions as a modal verb, and is followed by an infinitive form completing its meaning
- *dobbiamo* potentially highlights M5S members’ opinions and statements about what they should accomplish as part of a socio-political collective identity (Koller 2014: 148)
- collocation analysis to identify words strongly associated with *dobbiamo*, in a diachronic perspective
- collocation measure: MI - 6; span: 0L, 2R; frequency threshold: 5

## Pilot study 2: The modal verb form *dobbiamo* in diachrony (2)

	Collocate	MI	Coll. freq. (per 1,000 words)	Freq. in subcorpus (per 1,000 words)
2005	<i>pagare (to pay)</i>	9.70	1.25	4.37
	<i>anche (too, also)</i>	6.69	1.14	32.26
2013	<i>iniziare (to start)</i>	8.68	0.13	0.86
	<i>andare (to go)</i>	6.34	0.11	3.73
	<i>avere (to have)</i>	6.01	0.13	5.47
2018	<i>pensare (to think)</i>	6.85	0.12	1.99
	<i>intervenire (to intervene)</i>	6.50	0.06	1.40
	<i>tornare (to return/go back to ...)</i>	6.30	0.06	1.61

# Pilot study 3: A preliminary Keywords analysis (1)

Our aim is to look at the diachronic variation in our GriCo sampler

- statistical measure: Log Ratio (Hardie 2014)

$$LR = \log_2 \left( \frac{n_1/C_1}{n_2/C_2} \right)$$

- corpus tool: Lancsbox (Brezina, Timperley, McEnery 2018)
- default threshold (positive keywords  $s > 0.14$ , negative  $s > -0.15$ )

# Pilot study 3: A preliminary Keywords analysis (2)

We focused on *word types*, selecting the top 40 positive and top 40 negative keywords, not taking into account all the “**seasonal keywords**” (adapted from Gabrielatos & Baker 2008: 12), so those words related to specific event in the time period

Two keyword lists from the GriCo Sampler have been created- the newer one compared to the older one:

- GriCo sample 2018 vs GriCo Sample 2013
- GriCo sample 2013 vs GriCo Sample 2005

## 2018 vs 2013

## Keywords (+)

## Keywords (-)

Seasonal	Non Seasonal	Seasonal	No Seasonal
ceta [free trade agreement Canada-Italy]	blockchain	pdmenoelle	138 [law]
Conte	(class) action	capitan (captain)	comprimi (to compress)
efdd [political group in European parliament]	uninomiale (uninomial)	Findus (Birdseye)	espandi (to expand)
rousseau [web platform to vote]	paragone (comparison)	menoelle	pdmenoelle
Salvini	fake	Porcellum [a law]	menoelle
Raggi	potenziamento (improvement)	Sel [a party]	trasmesse (to broadcast)
Macron	pedaggi (toll)	Ligresti	Fo
jobs (a law)	startup	Bersani	copasir [a body of the Italian Parliament]
2018	City	Messora	rigor
Spazzacorrotti [a law]	urbana (urban area)	Renzie [mockery for Renzi]	pr
Daspo	force	Violante	Passaparola (word of mouth)
Virginia	giocattoli (toys)	Rodotà	pdexmenoelle
Gentiloni		Expo	Ventennio (two decades)
2019		Becchi	Assange
Juncker		Finocchiaro	Darfur
flat		Napolitano	Ineleggibilità (ineligibility)
that		Montis [mockery for Monti]	Cecità (blindness)
Casellati		Mussari	
wifi4eu		Shalabayeva	
Enrica		Van	
2017		Tares [ tax]	
Morandi		Carcano	
sindaca		Nuti	
l'azzardo		Paschi	
corrao		Zanon	
lab			
Appendino			
Bramini			

## 2013 vs 2005

## Keywords (+)

## Keywords (-)

Seasonal	Non Seasonal	Seasonal	Non Seasonal
m5s	Rimborsi (reimbursements)	Fiorani	C**o (as*)
pd	2012	Provera	Minerale (mineral)
stelle	Casaleggio	consorte (spouse)	vignetta (cartoon)
pdl	stabilità (stability)	romo (chromium)	vattene (leave!)
2013	segnalazione (advisory)	Tronchetti	polenta
pdmenoelle	Portavoce (spokesperson)	cab	<a href="http://www.repubblica.it">www.repubblica.it</a>
Napolitano	Attivisti (activists)	12100	fotovoltaico (photovoltaic)
Bersani	Pensioni (pensions)	c/c	svizzeri (Swiss)
2011	Movimento	voip	rom (Romani people)
Capitan (captain)	Liste (lists)	116276	Economist
Findus (Birds Eye)	emendamenti (amendments)	cin	Capitalia
Menoelle	Casta (caste)	ccrtit2t84a	Zuccherò (sugar)
Boldrini		swift	Confisca (requisition)
Porcellum [a law]		It35b050181210000000116276	omissis
Renzi		Benni	
Alfano		Ricucci	
Monti		FI	
Sel		titanio (titanium)	
Grecia (Greece)		dtc (digital Terrestrial Television)	
Expo		Petruccioli	
Matteo		MTP (media transfer protocol)	
Pizzarotti		13.10.05	
Facebook		Ruini	
Ligresti		Ruggiero	
Mps		Terry	
2008		Nichel	
Messora		Silicio (silicon)	
Cancellieri			

# Pilot study 3: A preliminary Keywords analysis (5)

- 2005: beginning of the movement (bank account information - fundraising campaign to support their mission - names of the politicians/political enemies - discussion of current political events)
- 2013: new use of language, established role in Italian politics (jokes/mockeries [monties and renzies] - typical Grillo's style, cf. Bortoluzzi & Semino 2016)
- 2018: focus on political strategies rather than others. Less protest, more discussion on politics (use of English words)

# Future developments

- **GriCo's potential**

- Large dataset
- Spans more than 10 years
- Includes users' comments
- Possibility to update

- **Variety of investigations**

- Diachronic and synchronic studies
- Language variation
- Political content (e.g. uncovering ideologies at the heart of a populist agenda)
- Identity building process - representation of collective identities and of 'the people' (see also Social Actor Representation framework, van Leeuwen 1996)
- Figurative language contributing to such representations
- Presences and absences in relation to major issues (e.g. the environment)
- Humour (based on Grillo's past as a comedian)

- **Platform for sharing**

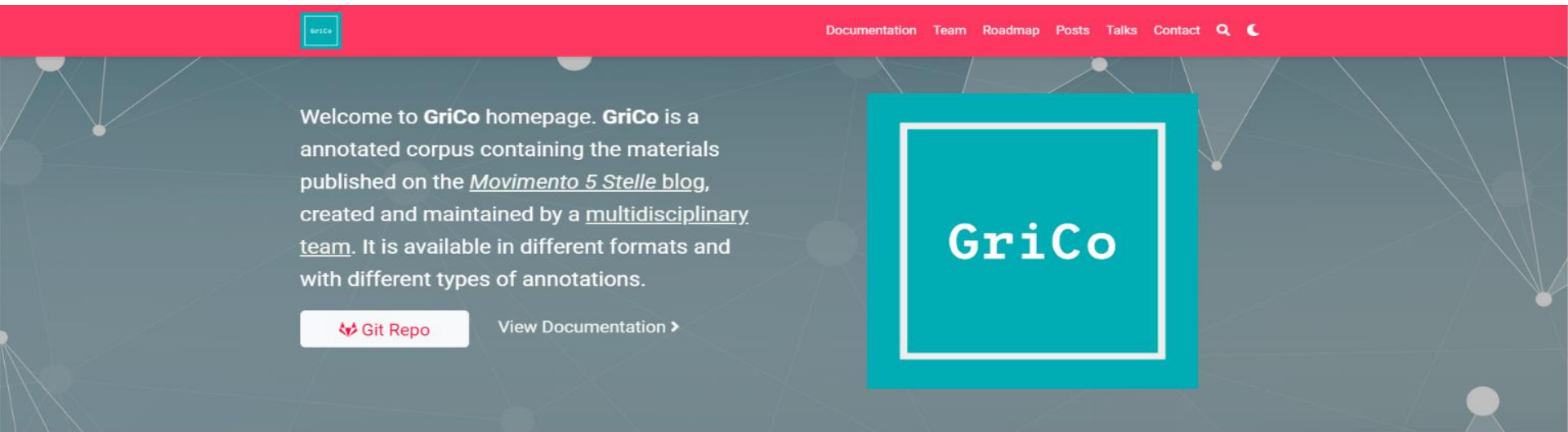




# Contacts

Email: [grico.project@gmail.com](mailto:grico.project@gmail.com)

Website: <https://grico.gitlab.io/>



The image shows a screenshot of the GriCo homepage. The page has a red header with a navigation menu containing links for Documentation, Team, Roadmap, Posts, Talks, and Contact, along with search and moon icons. A teal square logo with the text 'grico' is in the top left. The main content area has a dark grey background with a network diagram. A teal square with a white border contains the text 'GriCo'. Below the introductory text, there are two buttons: 'Git Repo' with a red Git icon and 'View Documentation >'.

Welcome to **GriCo** homepage. **GriCo** is a annotated corpus containing the materials published on the *Movimento 5 Stelle* blog, created and maintained by a multidisciplinary team. It is available in different formats and with different types of annotations.

[Git Repo](#) [View Documentation >](#)

# Bibliography

- Bentivegna, S. (2014). Beppe Grillo's Dramatic Incursion into the Twittersphere: Talking Politics in 140 Characters. *Contemporary Italian Politics*, 6(1), 73–88.
- Bortoluzzi, M., & Semino, E. (2016). "Face Attack in Italian Politics: Beppe Grillo's Insulting Epithets for Other Politicians". *Journal of Language Aggression and Conflict*, 4(2), 178–201.
- Diamanti, I. (2014). The 5 Star Movement: A Political Laboratory. *Contemporary Italian Politics*, 6(1), 4–15.
- Franzosi, P., Marone, F., & Salvati, E. (2015). Populism, and Euroscepticism in the Italian Five Star Movement. *The International Spectator*, 50(2), 109–124.
- Gabrielatos, C., & Baker, P. (2008). Fleeing, sneaking, flooding: A corpus analysis of discursive constructions of refugees and asylum seekers in the UK press, 1996-2005. *Journal of English linguistics*, 36(1), 5-38.
- Koller, V. (2014). Applying social cognition research to critical discourse studies: The case of collective identities. *Contemporary critical discourse studies*, 147-165.
- McEnery, A., & Baker, H. (2016). *Corpus linguistics and 17th-century prostitution: Computational linguistics and history*. Bloomsbury Publishing.
- Mosca, L. (2014). The Five Star Movement: Exception or Vanguard in Europe?. *The International Spectator*, 49(1), 36–52.

# Bibliography

- Musso, M., and Maccaferri, M. (2018). At the Origins of the Political Discourse of the 5-Star Movement (M5S): Internet, Direct Democracy and the 'Future of the past'. *Internet Histories*, 2(1-2), 98–120.
- Natale, S., & Ballatore, A. (2014). The Web Will Kill Them All: New Media, Digital Utopia, and Political Struggle in the Italian 5-Star. *Media, Culture & Society*, 36(1), 105–121.
- Stanyer, J., Salgado, S., & Strömbäck, J. (2016). Populist Actors as Communicators or Political Actors as Populist Communicators: Cross-National Findings and Perspectives. In T. Aalberg, F. Esser, C. Reinemann, J. Strömbäck, & C. H. de Vreese (Eds), *Populist Political Communication in Europe* (pp. 353–364). New York & London: Routledge.
- Taggart, P. (2000). *Populism*. Buckingham: Open University Press.
- van Leeuwen, T. (1996). The Representation of Social Actors. In C. R. Caldas-Coulthard and M. Coulthard (Eds), *Texts and Practices: Readings in Critical Discourse Analysis* (pp. 32–70). London: Routledge.
- Wodak, R. (2015). *The Politics of Fear: What Right-Wing Populist Discourses Mean*. London: SAGE.