Reducing Corruption through Legislative Action
A study of the Indian *Janlokpal Bill (2013)*

Devvart Poddar

December 16, 2016

Contents

1 Introduction

Corruption has become a major flash point in the Indian political debate post multiple scandals which rocked India in 2011. A major policy issue at the height of the debate was the creation of an independent corruption watchdog through legislative assent; the *Jan Lokpal Act* of 2013. The watchdog itself was recommended as far back as 1969, but was never created in the fluid political scenario.

There have been several studies which have looked at the impact of legislative actions in reducing corruption [See Quah, 2007, Prado et al. [2016]], however the results of the reforms are mixed. While Singapore was successful in the enforcement of their reforms in the city-state, similar changes in Brazil failed to reduce corruption. Moreover the study of corruption, be definition, is difficult to measure. Corruption is undertaken in the shadows of economic and political actions, and is notoriously difficult to predict. Watchdogs like the Transparency International (TI) use perception surveys which are conducted wig *experts* around the work. Other NGOs in India like the Center for Media Studies (CMS) also measure corruption through surveys, albeit with citizens across the country.

However, there are several issues which plague expert surveys. They are a subjective measure of perception which may be impacted by Individual biases. As the same individual is not studied across her lifetime, it is hard to dis aggregate the idiosyncratic biases from actual corruption. Thus I turn to the use of Indian media data sets to create an unbiased indicator of corruption.

2 Methodolgy

2.1 Media as Data

This study is not the first to use media as a source of data for economic research. Jansen and De Haan [2005] use media to identify the impact of communications from the European Central Bank (ECB) on exchange rate volatility. Similarly Baker et al. [2015] use media data to identify and measure economic policy uncertainty in 15 countries. However to the best of the author's knowledge, no study has used media as an indicator for corruption.

For the study, nearly 19000 articles were scrapped from Google News for a period of 8 years; from 2008 ~ 2016. The frequency of the articles is showcased in Figure 1 below. They identify two main changes in the approach of the study; *firstly* due to idiosyncrasies of Google News, there is a jump in the number of articles on corruption for the last month of the study. This is a clear outlive, possibly due to sorting on the basis of the time of the scraping. Those months will be ignored in the analysis.

Secondly we see a clear time trend and a seasonal component to the number of articles. While the time trend was expected, as the news media has evolved rapidly to growing digitization, the seasonal trend was not expected. Regardless moving forward, the study will take into account the time trend and seasonality in all further analysis.

2.2 Insights from Text

The media data sets are incredibly insightful in terms of the richness of the data provided. We can measure corruption up-till the states, and even cities to a certain extent. The wealth of data also allows analysis of the different aspects of corruption itself, i.e. the forms of corruption. We can also track the growth in corruption by different groups (corporate houses, politicians, NGOs etc) separately. However those are beyond the scope of this research. For the study, we will create a indicator of corruption using text, for the different states in India.

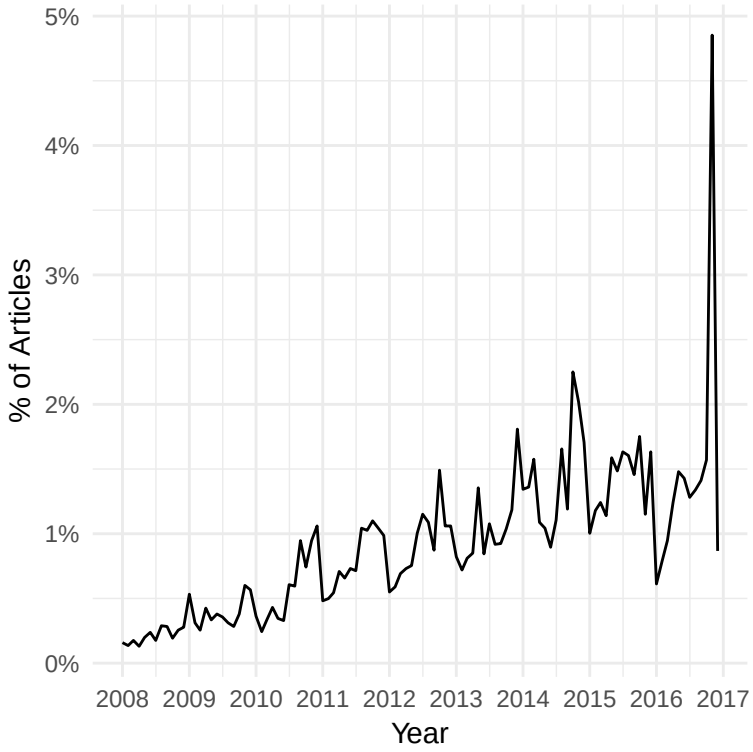

Figure 1: Trend in articles on corruption in India

2.2.1 Lemmatisation and Natural Language Programming

The first stage of the analysis involves the use of **Treetagger**, a tool developed by Schmid [1994] for annotating text with part-of-speech tags. **Treetagger** can annotate over 13 languages, and provides us with the *lemma*, the root form of a word. Lemmatisation is a powerful method of reducing the complexity of text, particularly in form of the different tenses, without changing the meaning of the text. The lemmatisation is further restricted on adjectives and adverbs, i.e. any word that qualifies / describes the context of the article is not modified. This would help in the succeeding stages to help create an index of corruption.

The text is also cleaned using a Natural Language Programming (NLP) framework [See Manning and Schütze, 1999, for a definative guide to NLP]. Natral Language Programming in an intersection of comupational-linuistics and artificial intelligence, aiming to make text *machine-readable*. As such, it is the base of all systems that depend upon understanding and analysing text (Siri from Apple is one of the best examples. The core component of Siri, understanding the *human* command, works through applications of NLP. See Dworetzky [2011]).

For the study, NLP is used primarly to detect *negation* and *emphasis*. Thus words like *not* and *no* which invert the meaning of a sentence are taken into account when building the index, as well as emphasis words like *very* and *greatly* which increase the emphasis of the sentence. Moreover the indicator is built at a sentence level, i.e we do not look at the entire text, but a window of words around corruption. This allows the index to identify the rising and falling trends of corruption in the different states.

To better understand the techniques of lemmatisation, and how it helps in creating the index, a small example is given below. The following quote is taken from a article on corruption against a prominent Indian politician;

> *One of India's most colourful and controversial politicians, Jayaram Jayalalitha, has been sentenced to jail for four years on corruption charges in a case that has lasted for 18 years. The chief minister of the southern state of Tamil Nadu was found guilty of amassing wealth of more than*

> *$10m (£6.1m) which was unaccounted for. She has to pay a 1bn rupee ($16m; £10m) fine and resign as chief minister.* (BBCNews [2014])

Upon cleaning and lemmatising, the text will change to as below;

> *one indias colourful controversial politician jayaram jayalalitha sentence jail four year corruption charge case last year chief minister southern state Tamil nadu find guilty amass wealth unaccounted pay bn rupee fine resign chief minister*

2.2.2 States and corruption

2.3 Regression Framework

3 Results

4 Conclusions

References

Scott R Baker, Nicholas Bloom, and Steven J Davis. Measuring economic policy uncertainty. Technical report, National Bureau of Economic Research, 2015.

BBCNews. Top india politician jayalalitha jailed for corruption, September 2014.

Tom Dworetzky. How siri works: iphone's 'brain' comes from natural language processing, stanford professors to teach free online course, November 2011.

David-Jan Jansen and Jakob De Haan. Talking heads: the effects of ecb statements on the euro–dollar exchange rate. *Journal of International Money and Finance*, 24(2):343–361, 2005.

Christopher D Manning and Hinrich Schütze. *Foundations of statistical natural language processing*, volume 999. MIT Press, 1999.

Mariana Mota Prado, Lindsey Carson, et al. Brazilian anti-corruption legislation and its enforcement: potential lessons for institutional design. *Journal of Self-Governance and Management Economics*, 4(1): 34–71, 2016.

Jon ST Quah. Combating corruption singapore-style: Lessons for other asian countries. *Maryland Series in Contemporary Asian Studies*, 2007(2):1, 2007.

Helmut Schmid. Treetagger. *TC project at the Institute for Computational Linguistics of the University of Stuttgart*, 1994.