Understanding the Linguistic Complexity of World Bank Annual Reports

Thea Madsen
t.madsen@exchange.hertie-school.org

Xiaoyan Hu
x.hu@mpp.hertie-school.org

Dang Ngoc Huy
h.dang@mpp.hertie-school.org

Abstract

Every year, the World Bank publishes an Annual Report that explores the challenges encountered by developing nations and what it is doing to help people help themselves. From its inception in 1944 until present day, its mission statement has shifted from the rebuilding of Europe towards fostering international development. In a way, these Annual Reports are mirrors that reflect the reality of their time and shine a light on the advancement of international development and the progress of developing nations. Yet, it is questionable if these reports could offer any real insights at all. Criticisms have been drawn from both within and outside of the organization on the unnecessarily complex and impenetrable language that is utilized in these reports, making them inaccessible to the very audience they are intended to serve. This paper, therefore, seeks to comprehend the linguistic sophistication of the World Bank Annual Reports by exploring the complexity of language used in these publications from 1947 up until 2018. By employing a combination of readability metrics, our research discovers that there existed a trend of increasing comprehensibility and decreasing language complexity from 1947 until 1956. However, from 1956 onward, this trend was reversed and the Annual Reports indeed became progressively more convoluted with a gradual but distinct decline in readability.

1. Methodology [1]

In *Bankspeak: The Language of World Bank Reports, 1946–2012*, Moretti and Pestre provided an intriguing investigation into the linguistic development of the notoriously difficult language of World Bank Annual Reports from 1946 until 2012. Whereas their study focused more on how clusters or words, specific terminologies and the Bank's vocabulary developed over the years in the Annual Reports,our research will take a more quantitative approach by investigating the textual and readability complexity of these reports in the aggregate level. We will explore lexical diversity and sophistication of these texts by endeavoring to answer our main research question: *Is there, in fact, a trend of increasing linguistic complexity in the World Bank Annual Reports?*

To identify the best approach on how to answer this, our research first explored different metrics that have been utilized in this line of research to capture readability and linguistic complexity in text data. Different political science research presents multiple diverging methods with which to measure complexity, one of which is making use of the metrics employed in education research. A pronounced example of this approach is that of Bischof and Senninger (2018) [3] which examined political party manifesto by applying LIX formula to generate a score on the readability of each document. The final scores are thereupon compared to those of another type of text ranging from Cinderella fairy tale to the philosophical writings of Habermas to locate the level of complexity in human perception.

There are certainly limitations posed by education research in analysing texts using those of another domains. Political science research such as Benoit et al. [2] have endeavored to resolve this issue by creating specialized metrics involving human coders as the "golden standards" to encapsulate prior knowledge and cognitive ability of human in understanding text as well as using a Bradley-Terry model to incorporate a broader range of textual complexity indicators to improve the accuracy in context specific text.

After surveying the available methods and approaches, our research decided to explore the complexity of the World Bank Annual Reports by employing 6 conventional readability formula from the field of education studies [6]:

1. Flesch Reading Ease Readability Formula (FRE)

$$206.835-(1.015*\frac{words}{sentences})-(84.6*\frac{syllables}{words}) \quad (1)$$

A classic readability formulae. This index is distinct from others in that it adopts a score range of decreasing

[1] Project GitHub page: https://github.com/huydang90/World-Bank-Corpus-Analysis

difficulty from 0 (hard to understand) to 100 (easy to understand). The score will operate as the evaluation of how easy it would be for a piece of text, and in this case, a specific Annual Report, to be understood and engaged with.

2. Flesch-Kincaid Readability Score

$$0.39 * \frac{words}{sentences} + 11.8 * \frac{syllables}{words} - 15.59 \quad (2)$$

This metrics was developed originally for the U.S. Navy but now it is considered as a highly suitable formula in education studies. It measures how difficult a reading passage is to understand and outputs the grade level someone must be at to easily grasp that passage of text.

3. Automated Readability Index (ARI):

$$4.71(\frac{characters}{words}) + 0.5(\frac{words}{sentences}) - 21.43 \quad (3)$$

ARI derives from two ratios: words difficulties and sentence difficulties. The final score indicates the age needed to understand the text according to American grade levels.

4. Coleman-Liau Index (CLI):

$$5.89 * \frac{characters}{words} - 30 * \frac{sentences}{words} - 15.8 \quad (4)$$

This metric does not account for the syllables forming the word but the length in characters. The final output is years of education necessary to grasp the text in question, similar to other metrics.

5. Gunning's Fog Intex (FOG)

$$0.4(\frac{words}{sentence} + \frac{complexwords}{words}) \quad (5)$$

This metric takes into account sentence length and words complexity. The obtained estimatation represent the number of years of education that a person requires to easily understand the text on the first reading. 5 is considered easy to understand, 10 is hard, 15 is more difficult, and 20 is higly complex.

6. Simple Measure of Gobbledygook (SMOG)

$$1.0430\sqrt{polysyllables\frac{30}{sentences}} + 3.1291 \quad (6)$$

This formula is designed to capture the understandability of a text. The result obtained from the formulas will be the the number of years in education that a person in the U.S would need to fully comprehend the text in question.

With these indexes, we will be able to measure the linguistic complexity of the economist language of these reports and understand whether they might cause readability problems. The textual sophistication measurement model,illustrated in Benoit et al. (2019) [2], though being a fascinating approach to understanding complexity, is difficult to replicate in this context, as it adds an extra layer of human coders and potentially biased judgement. In addition, the conclusion that Benoit et al. (2019) reached was that complexity analysis using FRE score is relatively similar to the best performing model of predicted probability illustrated in the paper, particularly in detecting trend and general direction. Therefore, we conclude that the FRE is still a good measure with which to gauge an understanding of sophistication in the World Bank Annual Reports.

2. Experiments

Data: All of World Bank Annual Reports from 1946 to 2018 are preserved in the organization's data portal. In total, there are 72 documents spanning the 72 years, each ranging from 60-200 pages in PDF format with accompanying graphs, tables and photographs. Rough OCR versions of these reports in plaintext format are also available, which will be the main data source that forms the corpus in our analysis. In the initial phase, we will process the data as is, without any further transformation.

Our main challenge with the available data was the flawed report files provided by the World Bank data portal: some of the reports were not available in plaintext format but only in scanned pdf version (i.e., Annual Report 2010 in English); others are defective text files with significantly reduced text that skewed the analysis of the language complexity (i.e., Annual Report 2007 which contains only 24 sentences compared to other documents that have roughly 500-5000 sentences depending on the years).

To resolve this issue, we set up our own OCR conversion by utilizing the Tesseract library. The scanned reports were first converted into images, which were subsequently iterated through by the text recognition function from Tesseract. With this method, our project was able to obtain the desired plaintext format for these report files, which are surprisingly of higher quality than the version available on the official World Bank data portal, in terms of the identified text.

Evaluation method: As aforementioned in the Methodology, our main metrics will be: FRE, Automated Readability Index, Coleman Liau, Flesch Kincaid, Gunning Fog Index, and SMOG. FRE score will measure the ease of readability (the higher the score, the easier it is to peruse and vice versa). The rest of the metrics capture the readers' obtained grade level necessary to understand the text in question (the higher the grade, the more difficult it is to peruse).

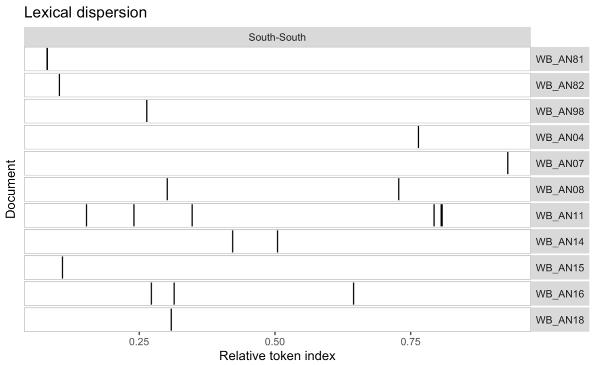

Figure 1. South-South occurrence over time

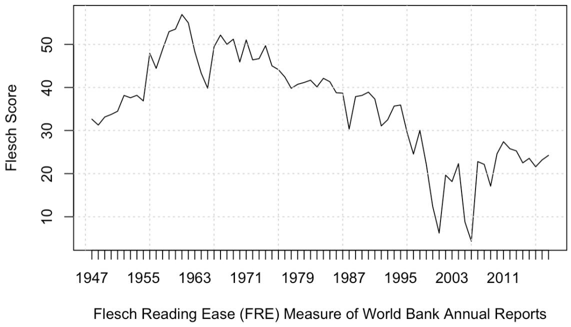

Figure 2. FRE score of World Bank Annual Reports 1947-2018

Experimental details: In our initial experiments, due to the flawed data from certain years, the analysis produced unreliable results with multiple outliers in document length, sentences and tokens. However, after obtaining the fixed data from our OCR conversion, the analysis was able to produce much more consistent results with regards to these factors.

Preliminary exploratory analysis on batches of selected World Bank Annual Reports revealed tendencies in language complexity that, indeed, satisfy our initial hypothesis that the discourse in these reports became more complicated over time. FRE score pinpoints that the 1947-96 report scores fall between the values of 30 and high 50s meaning that the reports would be understandable to most people over the age of about 15 years but below university level; after 1996, FRE scores decrease as the reports became much more difficult to understand for people below university education. This finding is in line with our hypothesis that the language of the reports over time develops to become more difficult to understand by the general public. Exploratory analysis with Flesch-Kincaid and Gunning Fog measure also revealed similar results with initial scores indicating academic-level papers gradually transforming into much more complex texts.

Graphing the lexical dispersion allowed us to track the development of technocratic or complex words, such as South-South cooperation, that might pose difficulty for wide audiences to comprehend (Figure 1). The analysis revealed that words began appearing sporadically in the reports from the early 1980s and then every year since 2014 except for 2017 testifying to the increased complexity in the report discourse. This line of analysis could present an interesting avenue for further research exploration on the complexity at the sentence and bigrams levels.

Collocation analysis also allows for the identification of multi-word expressions frequently used in the corpus, which are likely to further complicate the reports. This shows that the most frequently used collocations are, unsurprisingly: *World Bank* and *United States* but, then, interestingly, *Executive Directors, East Asia, Financial Statements* and *South Asia*, which shows a top-down approach of the reports to development aid from the donor countries headed by the World Bank and USA but also a surprising focus on Asia.

For our final analysis, by applying the readability indexes from the *textstat_readability* package in R, our project was able to capture the majority of the metrics intended for analysis. The main challenge was the runtime as the corpus is quite large and thus exerts a heavy toll on computing power.

Furthermore, as each metrics has a different method in calculating the readability of the language, there is bound to be variations that might interfere with the interpretation of their scores. To account for this issue, the different metrics for analyzing grade level needed to understand the text data were combined together and their averages over time were recorded with the goal to untangle the main signal in understanding the complexity, while reducing the noise.

Results and interpretations: From both the individual metrics calculations and the combined scores (Figure 1 2), our research was able to identify two general trends in the readability of the World Bank's Annual Reports: from 1947 until 1954, the language complexity level was stable but slightly decreasing over time; from 1955, there was a sudden and significant drop of complexity, which subsequently began to rise continuously with two additional sudden forward leaps in levels in 1964 and 1995. This finding upholds our hypothesis that there is indeed a trend of increasing complexity in the World Bank Annual Reports over time.

By studying the Bank's chronology and historical development, our research was able to match some potential explanation with these jumps in language complexity level.

A possible reason for the first sudden discontinuity - a pronounced drop in complexity in 1955 - could be that the year 1956 saw the establishment of the International Finance Corporation - World Bank's member that focuses on investment lending for private companies and financial institutions in developing countries. Essentially, this was

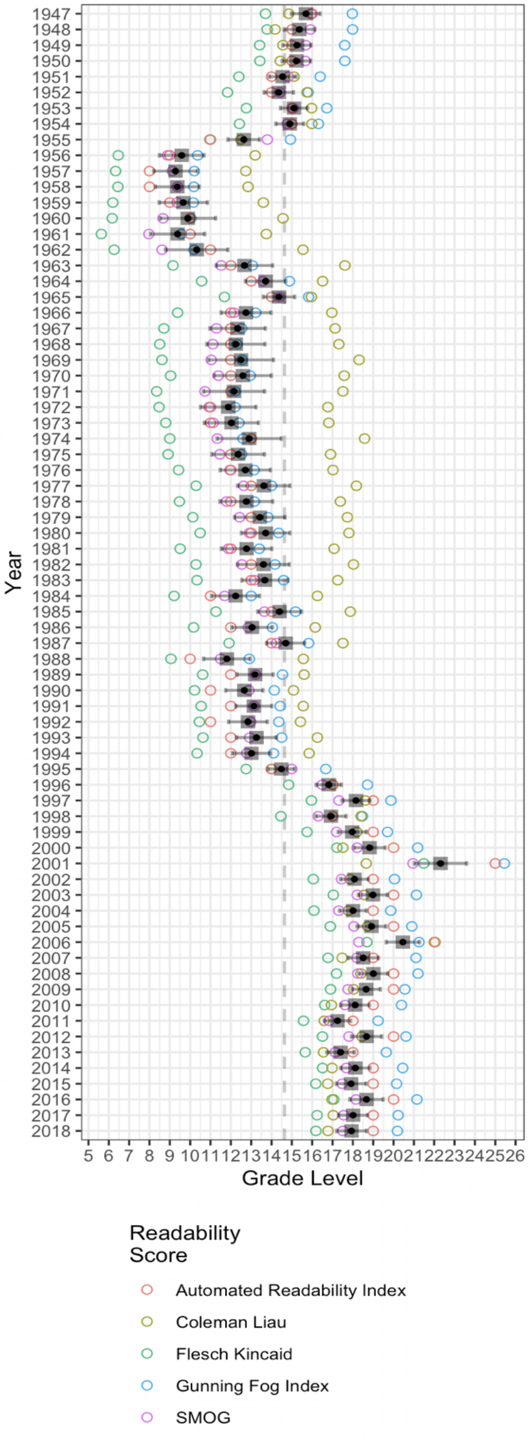

Figure 3. Readability of World Bank Annual Reports 1947-2018

the demarcation year when the World Bank began to shift its mission from reconstruction of Europe towards international development. We conjecture that this restructuring of the Bank and the change in its clientele and stakeholders base towards developing nations were the main motivation for the shift in language complexity towards simplicity and readability in the World Bank reports. If this was the case, it seems that the Bank itself was aware of the importance and impact of having the language of its reports be approachable to its audience. However, as we can witness in the chart, this would not always be the case as the reports' readability would deteriorate over the years, and skewered towards the indecipherable.

For the second period of sharp rise in complexity from 1963-1965, we were not able to pinpoint any specific event that might reflect this change. However, delving deeper into the individual document level, it seems that this period was considered as one of the most active in the history of the Bank and its affiliate institutions, with vast amount of loans, credits, and other commitments delivered around the world. This expansion in activities and operations might partially explain for the spike of linguistic complexity in the reports.

As for the third discontinuity, the year 1995 witnessed the ascension of James Wolfensohn to the office of president of the World Bank. A tough, unconventional, and no-nonsense leader, Wolfensohn ushered in an era of swift reform against the Bank's stagnant bureaucracy and revitalized a sense of purpose and direction for its mission [5]. He demanded the best of his people and expected ideas and work of the highest calibre. This drastic change in leadership style and expectation of results might have contributed to the further codification and standardization of World Bank's complex technical language, as a representation of the institution's renewed drive towards excellence.

3. Future work

After confirming our initial hypothesis, we plan to deepen our research by exploring the following avenues:

- Compare the readability of the World Bank Annual Reports with another publication from the Bank such as the World Development Report, which focuses on the institution's annual and in-depth analysis of the economic, social, and environmental state around the globe. The comparison will help to inform whether this trend of increasing linguistic complexity is an institutionalized phenomenon across different publications and affiliate members of the World Bank, or an isolated trademark of the Annual Reports;

- Compare the readability of the Annual Reports from World Bank with those of another international financial institutions such as the African Development Bank or the Asian Development Bank to investigate whether this trend is universally shared among the counterparts of the development world;

- Explore the language complexity in the sentence, words and bigrams level.

References

[1] Franco Moretti and Dominique Pestre. *Bankspeaks: the Language of the World Bank Report*. New Left Review, vol. 92, 2015, pp. 75–99, 2015

[2] Kenneth Benoit, Kevin Munger, and Arthur Spirling. *Measuring and Explaining Political Sophistication Through Textual Complexity*. American Journal of Political Science, 2019.

[3] Daniel Bischof and Roman Senninger, *Simple Politics for the People? Complexity in Campaign Messages and Political Knowledge*. European Journal of Political Research 57 (2): 473–495, 2018.

[4] World Bank Group Archives, *World Bank Group Historical Chronology*.

[5] Sebastian Mallaby, *The World's Banker: A Story of Failed States, Financial Crises, and the Wealth and Poverty of Nations*. Council on Foreign Relations Books (Penguin Press), 2006

[6] Thomas Jakobsen and Thomas Skardal, *Readability Index*. Agder University, Faculty of Engineering and Science, 2007.

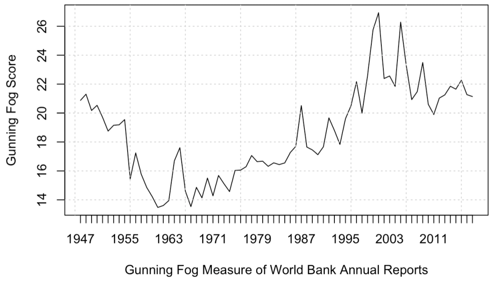

Figure 4. Gunning Fog score of World Bank Annual Reports 1947-2018

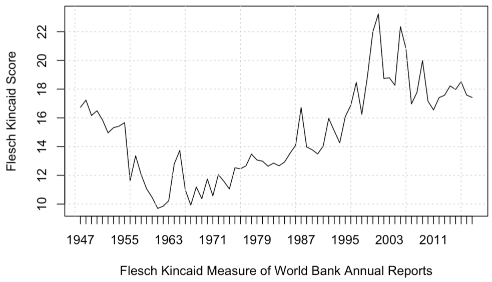

Figure 5. Flesch Kincaid score of World Bank Annual Reports 1947-2018