Thea discussion thoughts Meeting 19.10.

**Points**

1. What to do with outliers
   1. Huy worked with the tesseract and found that the outliers don’t matter too much:



* 1. In this code we can add measures together and weight them to take out the noise
  2. Big changes from 1956 in writing of the reports; after this steady increasing complexity of the writing of the reports
  3. Might have to do manual conversions of the years that we suspect of having errors in the conversion
  4. Between the years, there might be minor variations e.g. having different people write the reports

1. Readability measures
   1. Thea suggestions (please see notes on GitHub):
      1. The Flesch-Kincaid grade level
      2. The Automated Readability Index - ARI –
      3. The FORCAST formula
      4. The Linsear-Write calculation
2. Agree on pre-processing & Lexical diversity
   1. Thea suggested code uploaded on GitHub with preliminary interpretation comments
   2. Show consideration of the variations between the reports and take methodological steps to counter bias
3. Improvements to Slava's feedback for the proposal
   1. Adele suggested looking into readers pre-existing understanding and knowledge
      1. Suggestion for perspectivation for e.g. validity discussion e.g. human coding as Huy mentioned
      2. Also expanding the investigation to comparing with other kinds of reports e.g. the development reports or IFC, African development reports etc.
      3. Point of comparison would be to look at if the increasing complexity is tied to the World Bank - e.g. an institutionalized problem within the World Bank - or to the discourse in development aid - e.g. is the growing complexity a problem only for the World Bank?
      4. Download the papers online and convert them

**Division of work**

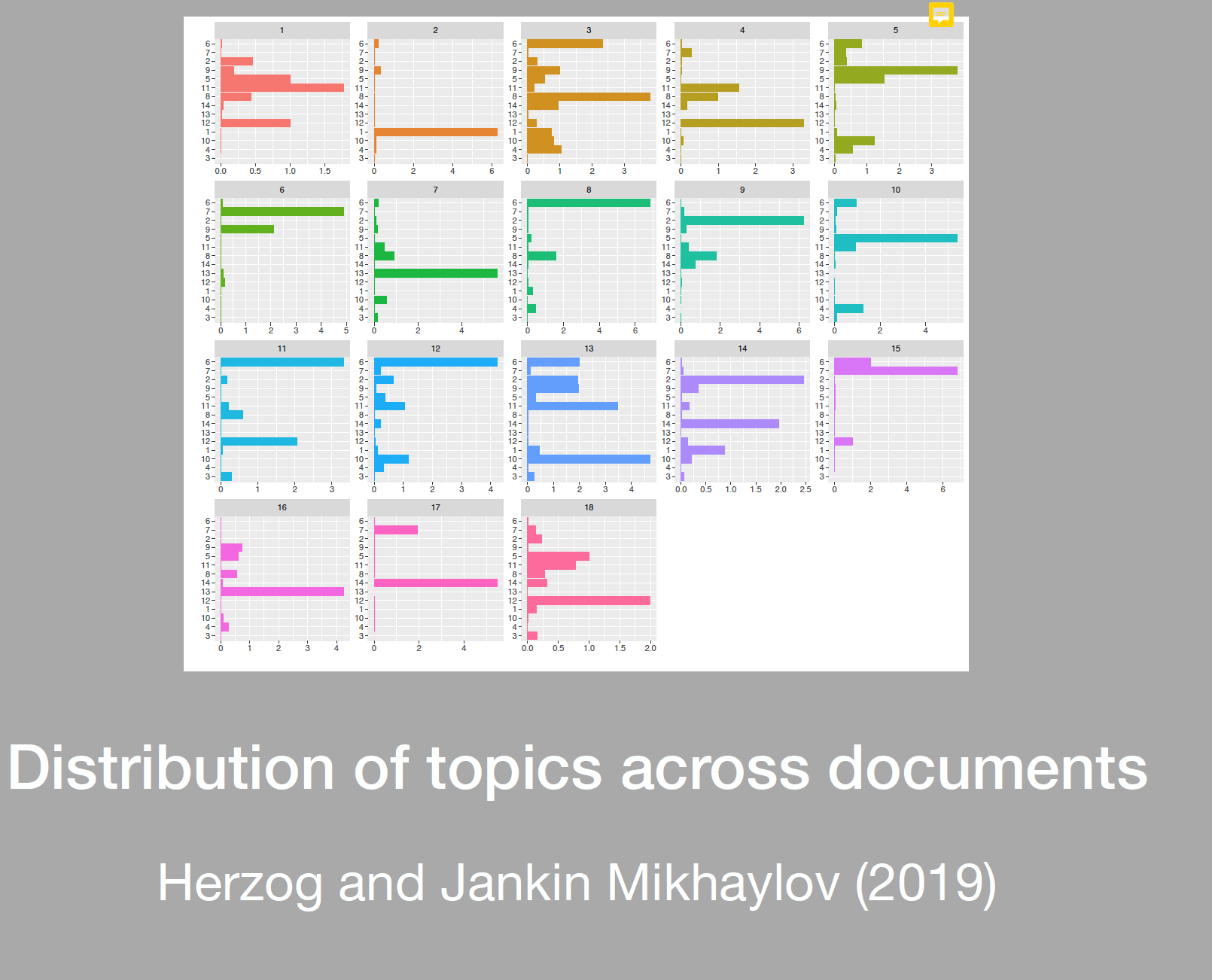
1. Short literature review (Adele)
   1. Reuse some of the text from the proposal
      1. Stanford paper and the paper about the US Presidential speech or another one that is closer to financial texts
2. Pre-processing of the data and exploratory analysis
   1. Take out
      1. Stop words, numbers, punctuations
      2. Don’t convert words to lower case to make it clearer for us what the words are
   2. Lexical diversity
      1. Word clouds
      2. Lexical diversity across the years
3. Matrix (Adele)
   1. talk about the matrix we have chosen and why
      1. What is the implication
      2. Explain the different mathematical equations
   2. FRE score and the scores in the graph (6 in total)
4. Experiments
   1. Write about the data problems in conversion
   2. Experimental details – what we run and enter graphs

Finish everything before 18 Sunday 20.10.

* Have a look and talk together

**Maybe in final report look at**

1. “key” words: words selected because of special attributes, meanings, or rates of occurrence
2. What about this method for uncovering the topics in the different reports (See endnote with screenshot attached)?[[1]](#endnote-1)

1. Screen shot from Lecture 1, slide 24 in PDF.  [↑](#endnote-ref-1)