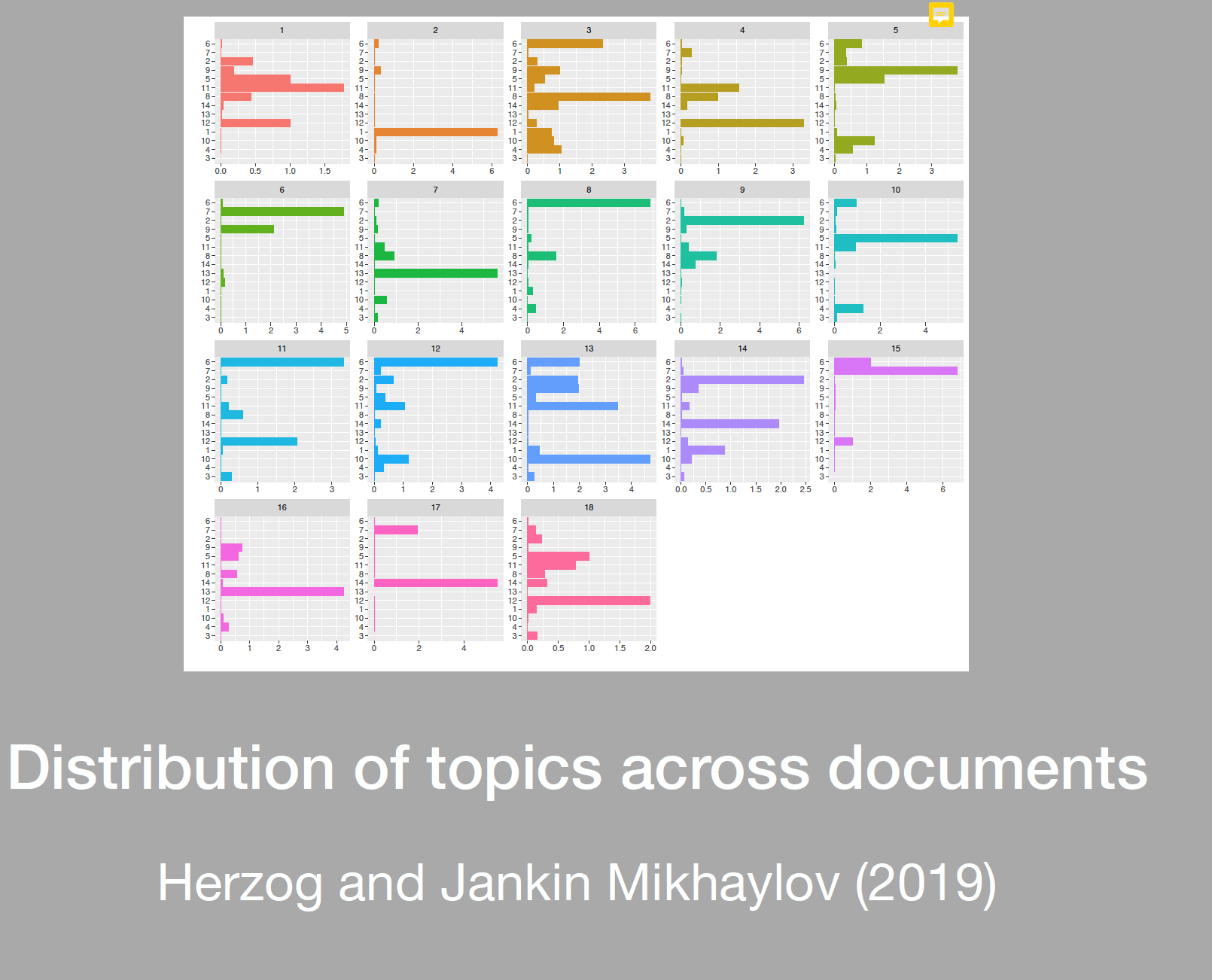
Thea discussion thoughts Meeting 19.10.

**Points**

1. What to do with outliers
2. 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. Question: Relevant to look at Euclidean distance to compare documents?
3. Agree on pre-processing & Lexical diversity
   1. Thea suggested code uploaded on GitHub with preliminary interpretation comments
4. Improvements to Slava's feedback for the proposal

**Further questions - Do we want to change anything in regards to our:**

1. How can we check the corpus length so we ensure that the corpus are actually comparable?
2. Unit of analysis (document, paragraph, sentence, etc.)
3. Features including tokens, equivalence classes of tokens (dictionaries), selected phrases, human-coded segments (of possibly variable length), linguistic features etc.
   1. I think this is very important because I have a hypothesis that the reports develop from using simpler ‘stand-alone’ words to using expressions and phrases - how can we pick up on this in defining our features as with the collocations?
   2. Identifying collocations
4. “key” words: words selected because of special attributes, meanings, or rates of occurrence
5. What about this method for uncovering the topics in the different reports (See endnote with screenshot attached)?[[1]](#endnote-1)
6. We are currently investigating the readability of the reports - providing an estimate based off the text word length, syllable length, etc. - but what about looking into the complexity - e.g. if a word contains three syllables or more - or the lexical diversity measuring how many types occur per fixed word rate (a normalised vocabulary measure)?

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