Topic modeling of Literacy Education articles: five decades of scholarship

LING 2340- Data Science for Linguists

Gianina Morales

University of Pittsburgh

Introduction

- Topic modeling is a data mining and machine learning technique that automatically analyzes texts to identify latent topic structures. The most widely used model in humanities is Latent Dirichlet Allocation—LDA.
- I use topic modeling to analyze the trends over time in literacy research and scholarship in one leading journal and a conference papers journal in the field of Literacy education, both from the same disciplinary association.
- My research questions are:
    - What are the trends in topics of literacy education research and scholarship over more than five decades (1969-2022) of the focal journals?
    - How do the topics have changed over time?

Relevant literature

- I have reviewed previous studies that used topic modeling to observe trends in research over time. For example, Wang et al. (2017) applied topic modeling to review trends in the literature on educational leadership.
- Other relevant studies are content analysis developed in the field of literacy education (e.g., Baldwin et al., 1992; Guthrie et al., 1983; Parsons et al., 2016). These studies identified topics and changes in the field over time. For example, Baldwin and colleagues identified a change in the focus of the Literacy Research Association from being interested in college reading in the 1950s to turning toward school children in the 1990s.

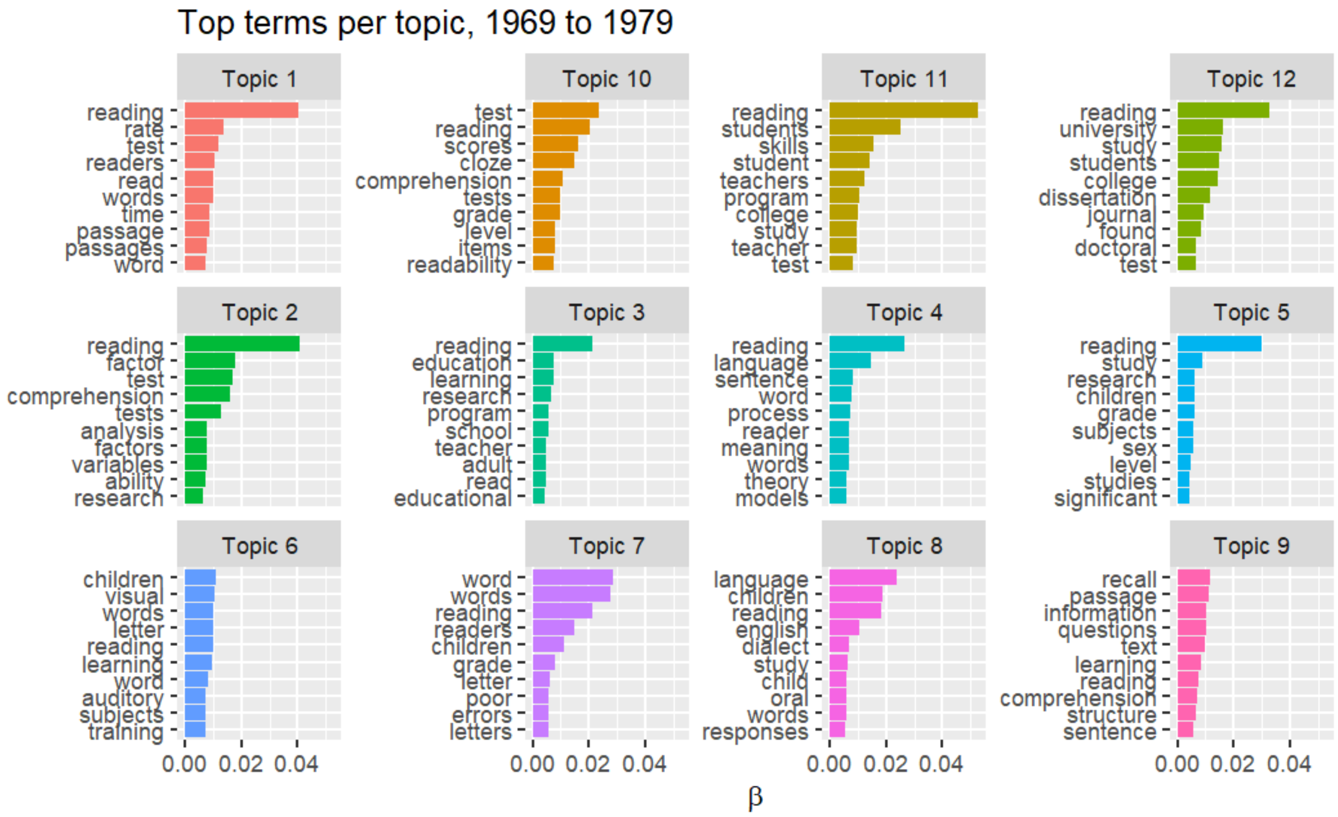
Method

- I followed the book “Text Mining with R: A Tidy Approach” (Silge and Robinson, 2022) to apply the methodology of topic modeling. This had some limitations associated with the packages used.
- Main packages: “tidytext” and “topicmodels.”

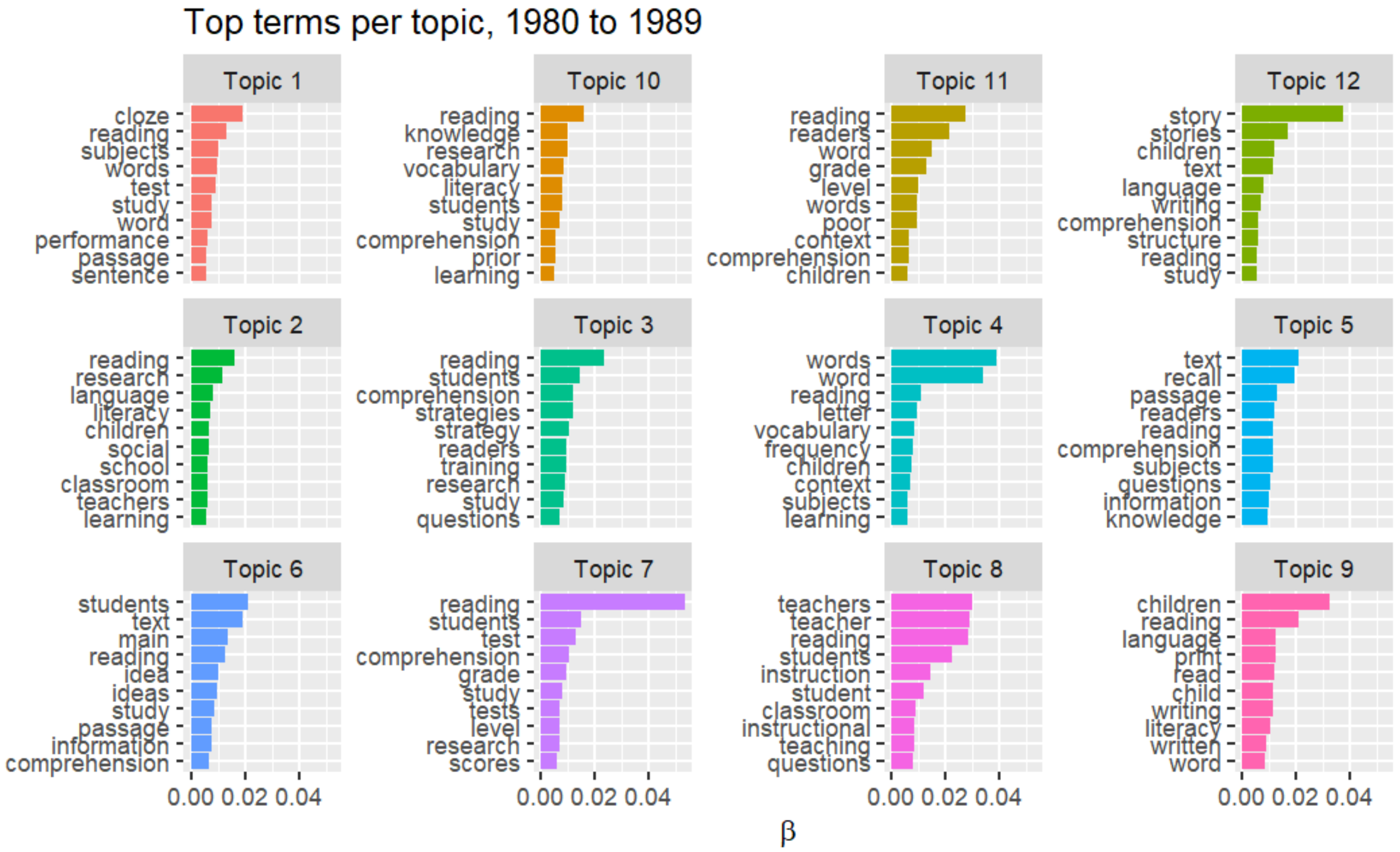
Data

- Part of a research project (confidential agreement)
- 3,131 articles published between 1969 and 2022 in two literacy journals (a research journal and a conference journal) of a leading association in the field.
- After pre-processing, I obtained 9,134,631 tokens. In the process of tokenization, I consider stop-words from the lexicon “Onix” and a list that I created based on conversations about stop-words on GitHub.
- The resulting data frame was so big that my computer did not have the capacity to model it. I ended up executing the topic models method by decade.
- After estimating 10 topics by decade, I manually analyzed the terms by topic to find a theme in each one. Then I plot the themes in search of trends.

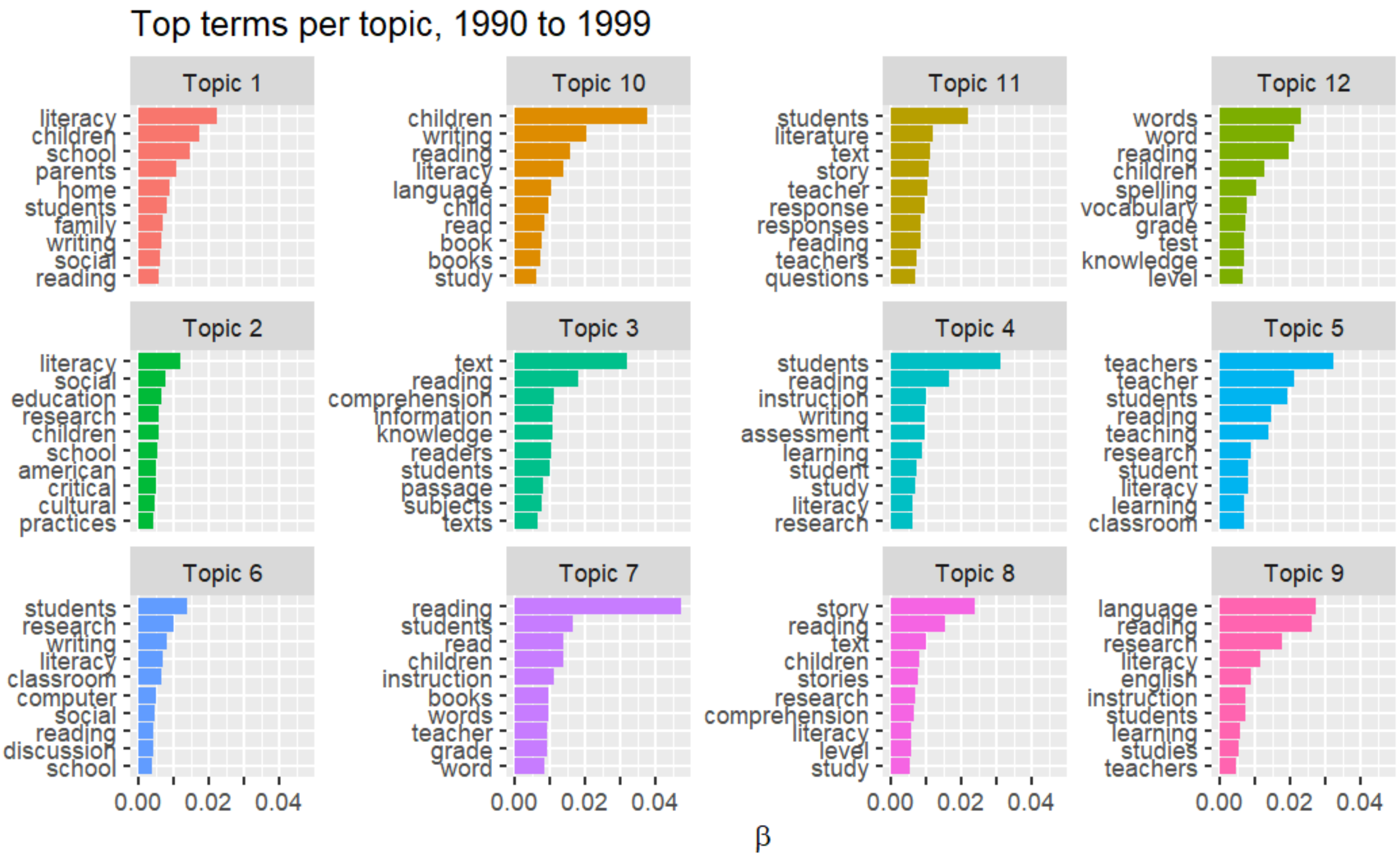
Analysis RQ1– topics by decade 1969-1979



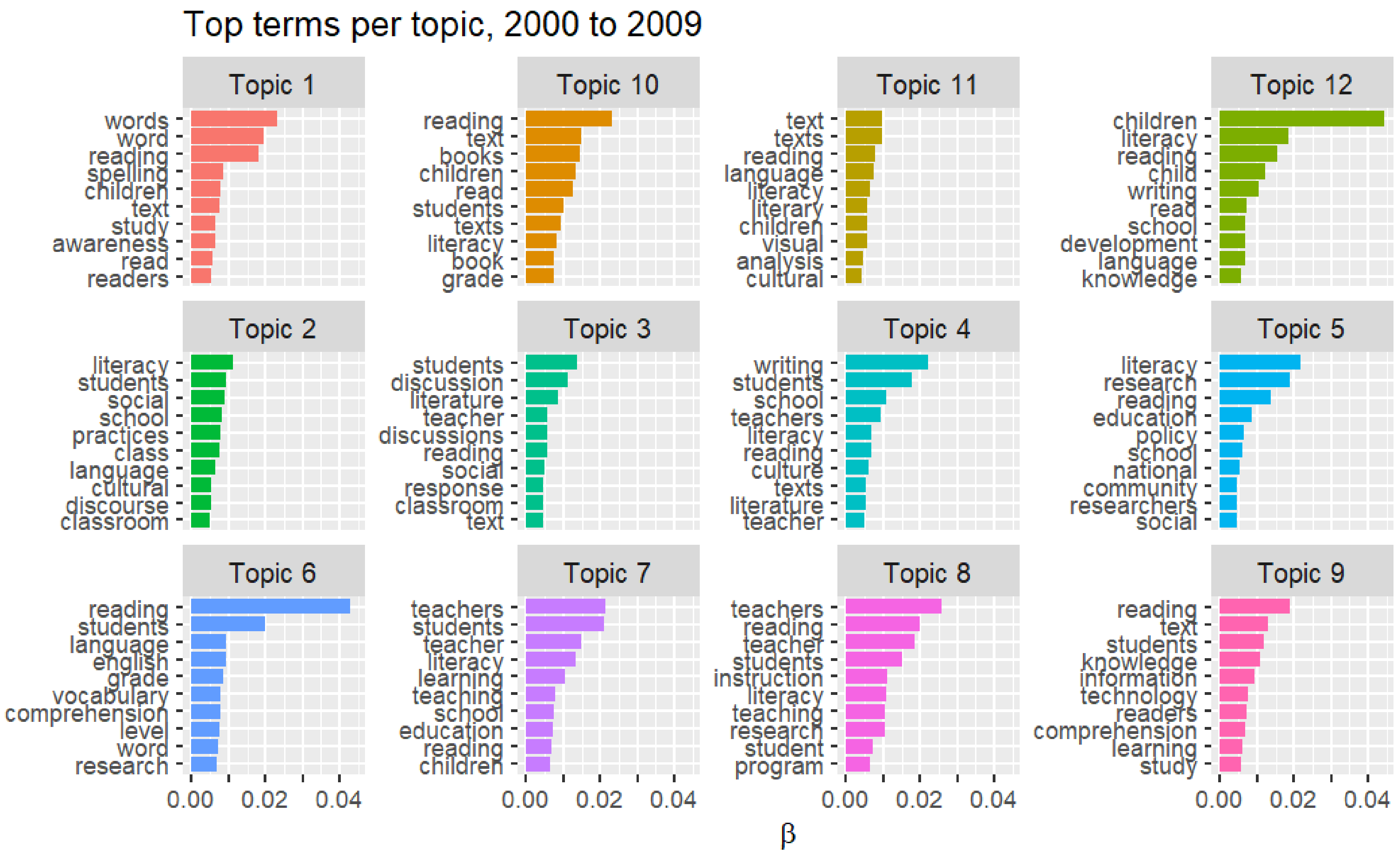
Analysis RQ1– topics by decade 1980-1989



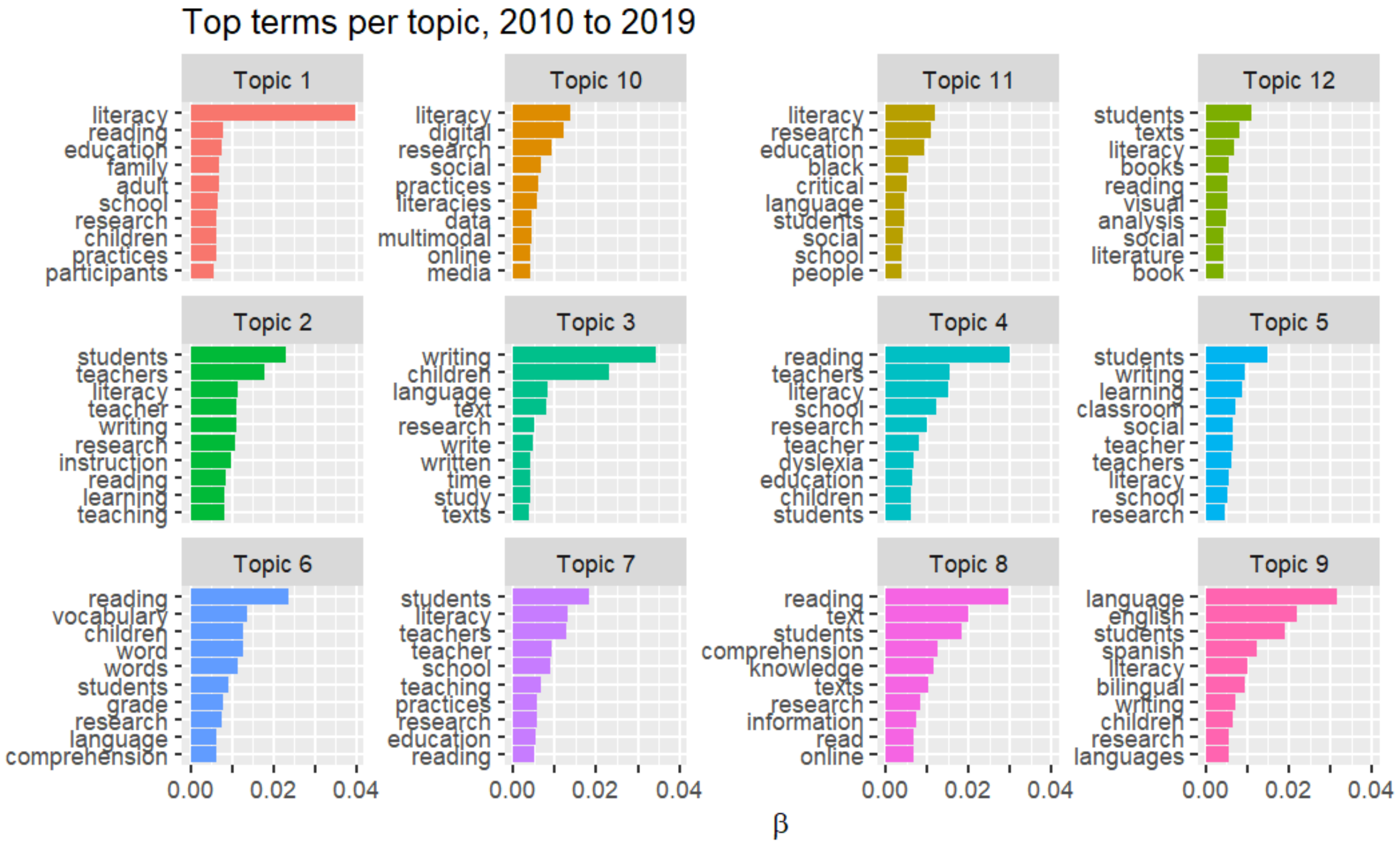
Analysis RQ1 – topics by decade 1990-1999



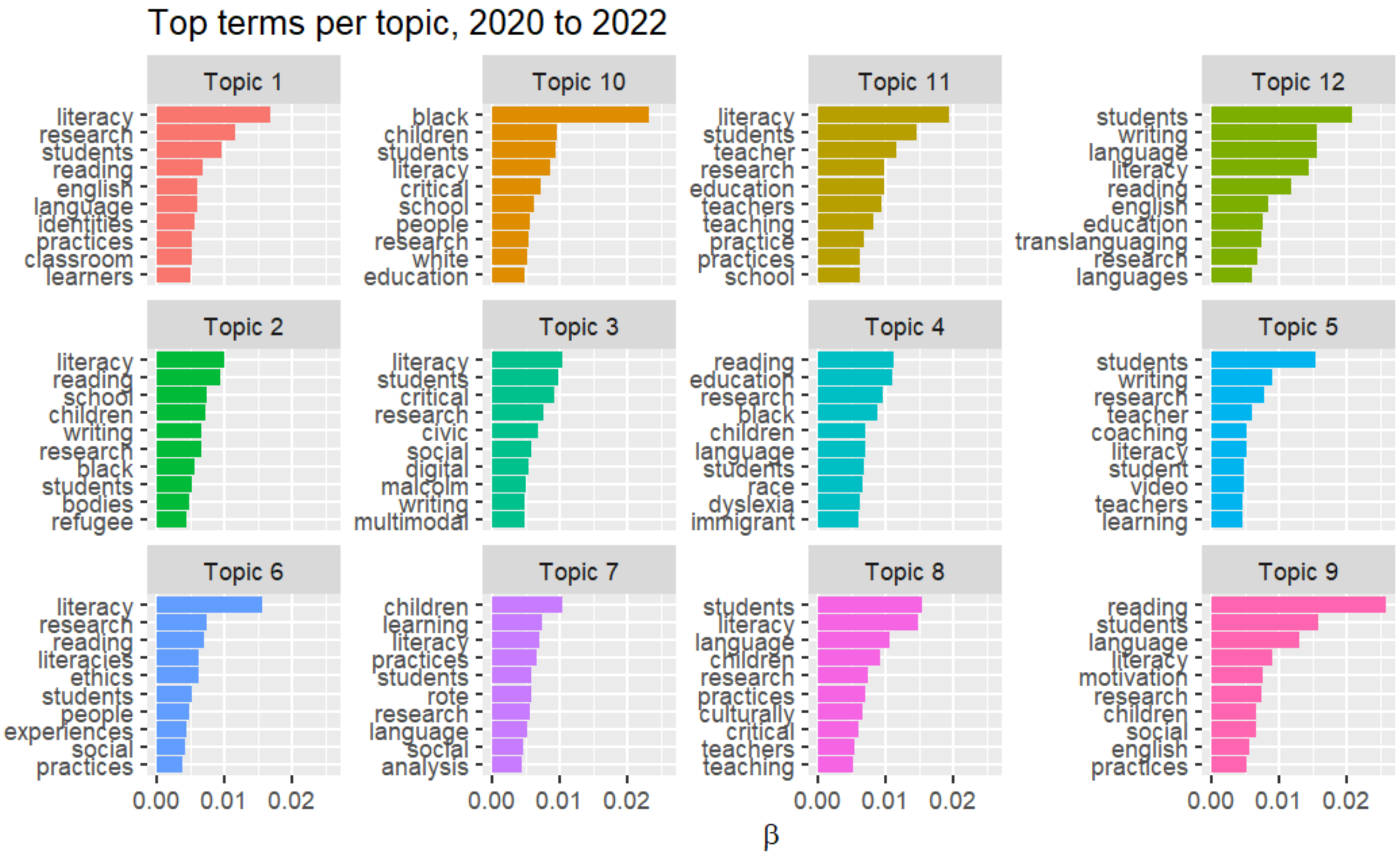
Analysis RQ1– topics by decade 2000-2009



Analysis RQ1– topics by decade 2010-2019



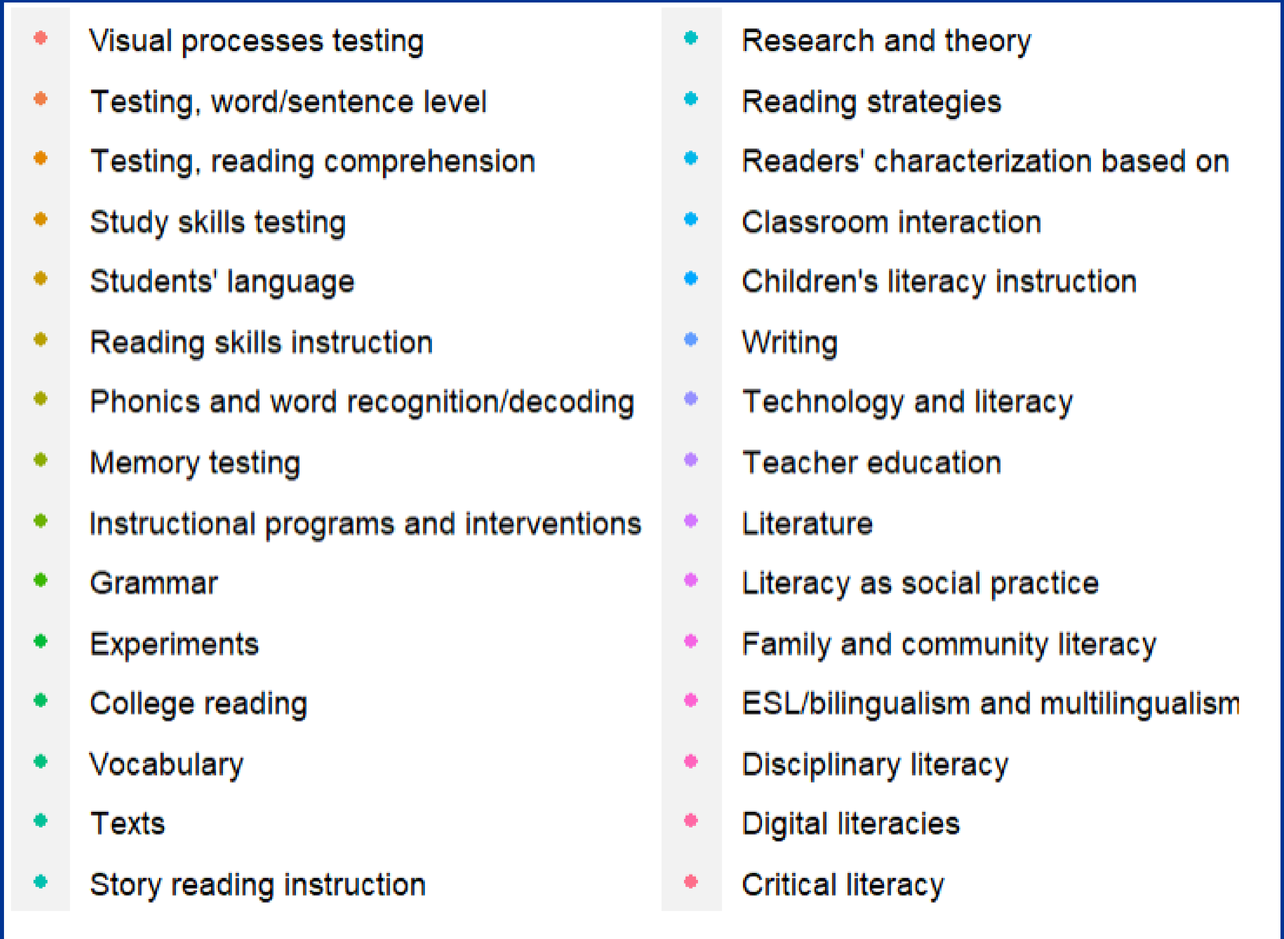
Analysis RQ1– topics by decade 2020-2022



Top terms per topic, 2020 to 2022

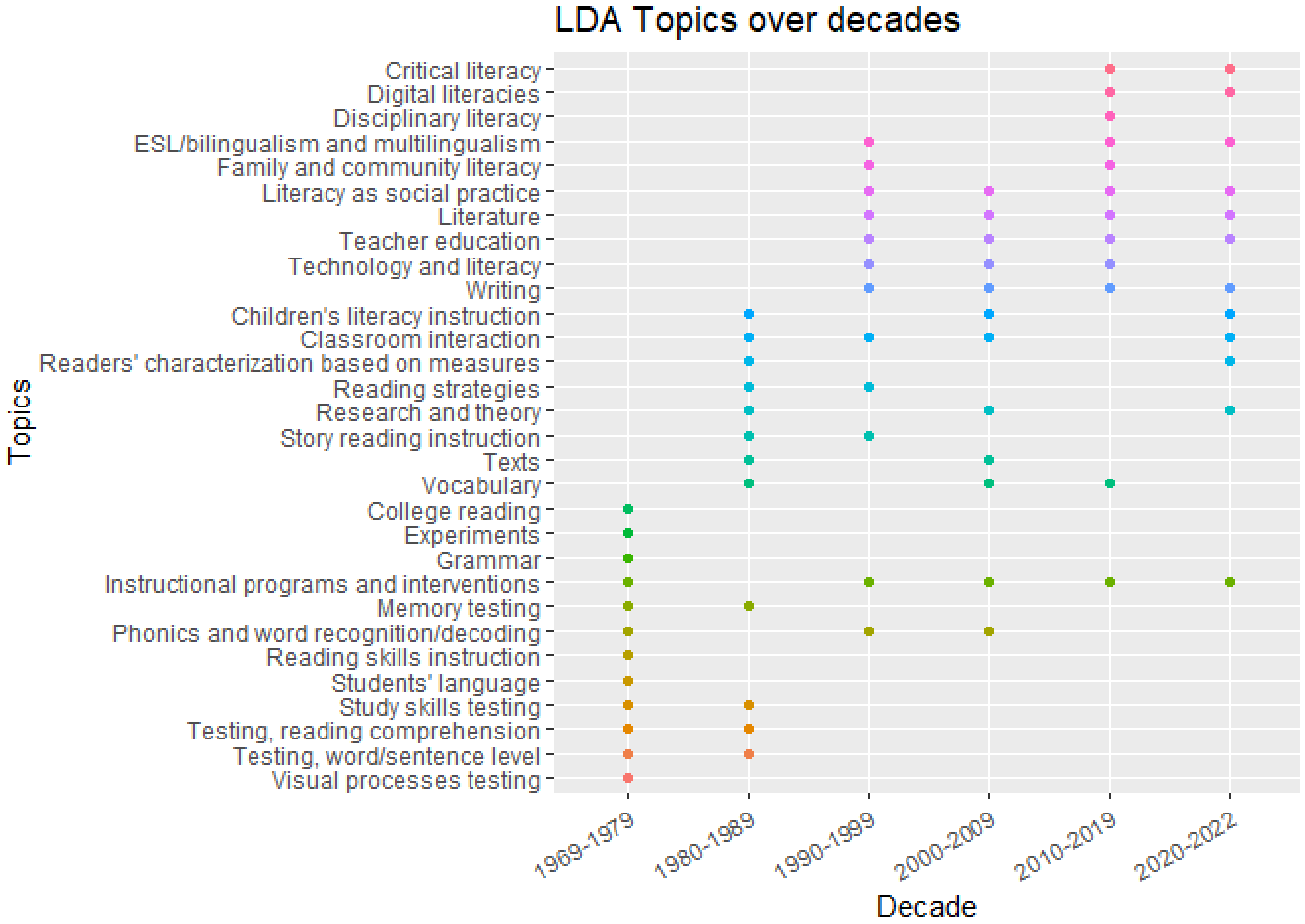
University of Pittsburgh

Topics manually coded



- ● Visual processes testing
- ● Testing, word/sentence level
- ● Testing, reading comprehension
- ● Study skills testing
- ● Students' language
- ● Reading skills instruction
- ● Phonics and word recognition/decoding
- ● Memory testing
- ● Instructional programs and interventions
- ● Grammar
- ● Experiments
- ● College reading
- ● Vocabulary
- ● Texts
- ● Story reading instruction
- ● Research and theory
- ● Reading strategies
- ● Readers' characterization based on
- ● Classroom interaction
- ● Children's literacy instruction
- ● Writing
- ● Technology and literacy
- ● Teacher education
- ● Literature
- ● Literacy as social practice
- ● Family and community literacy
- ● ESL/bilingualism and multilingualism
- ● Disciplinary literacy
- ● Digital literacies
- ● Critical literacy

Analysis RQ2– trend in topics



LDA Topics over decades

Discussion and future work

- Apparently, there are some changes in the topics over time. Some of them disappear over time, while others appear.
- My results are similar to what the literature said about the field of literacy education scholarship.
- My analysis is limited by the human part of the analysis. I am searching for options of equipment to be able to run the code with the whole corpus.
- Lately, I discover that the method of my reference book was not enough for my intentions, so I want to study other R packages in the future.

References

- Baldwin, R. S., Readence, J. E., Schumm, J. S., Konopak, J. P., Konopak, B. C., & Klingner, J. K. (1992). Forty Years of NRC Publications: 1952–1991. Journal of Reading Behavior, 24(4), 505–532. https://doi.org/10.1080/10862969209547793
- Guthrie, J. T., Seifert, M., & Mosberg, L. (1983). Research Synthesis in Reading: Topics, Audiences, and Citation Rates. Reading Research Quarterly, 19(1), 16. https://doi.org/10.2307/747334
- Parsons, S. A., Gallagher, M. A., & the George Mason University Content Analysis Team. (2016). A Content Analysis of Nine Literacy Journals, 2009-2014. Journal of Literacy Research, 48(4), 476–502. https://doi.org/10.1177/1086296X16680053
- Silge, J. and Robinson, D. (2022). Text Mining with R. https://www.tidytextmining.com/
- Wang, Y., Bowers, A. J., & Fikis, D. J. (2017). Automated Text Data Mining Analysis of Five Decades of Educational Leadership Research Literature: Probabilistic Topic Modeling of EAQ Articles From 1965 to 2014. Educational Administration Quarterly, 53(2), 289–323. https://doi.org/10.1177/0013161X16660585

Thank you for listening!