WRR in Latin America - 11/01/2018

Update on the Download Party

Hervé Guillon*

UC Davis - LAWR

* *hguillon@ucdavis.edu*

1. Methods

2. Results

3. Conclusion and next steps

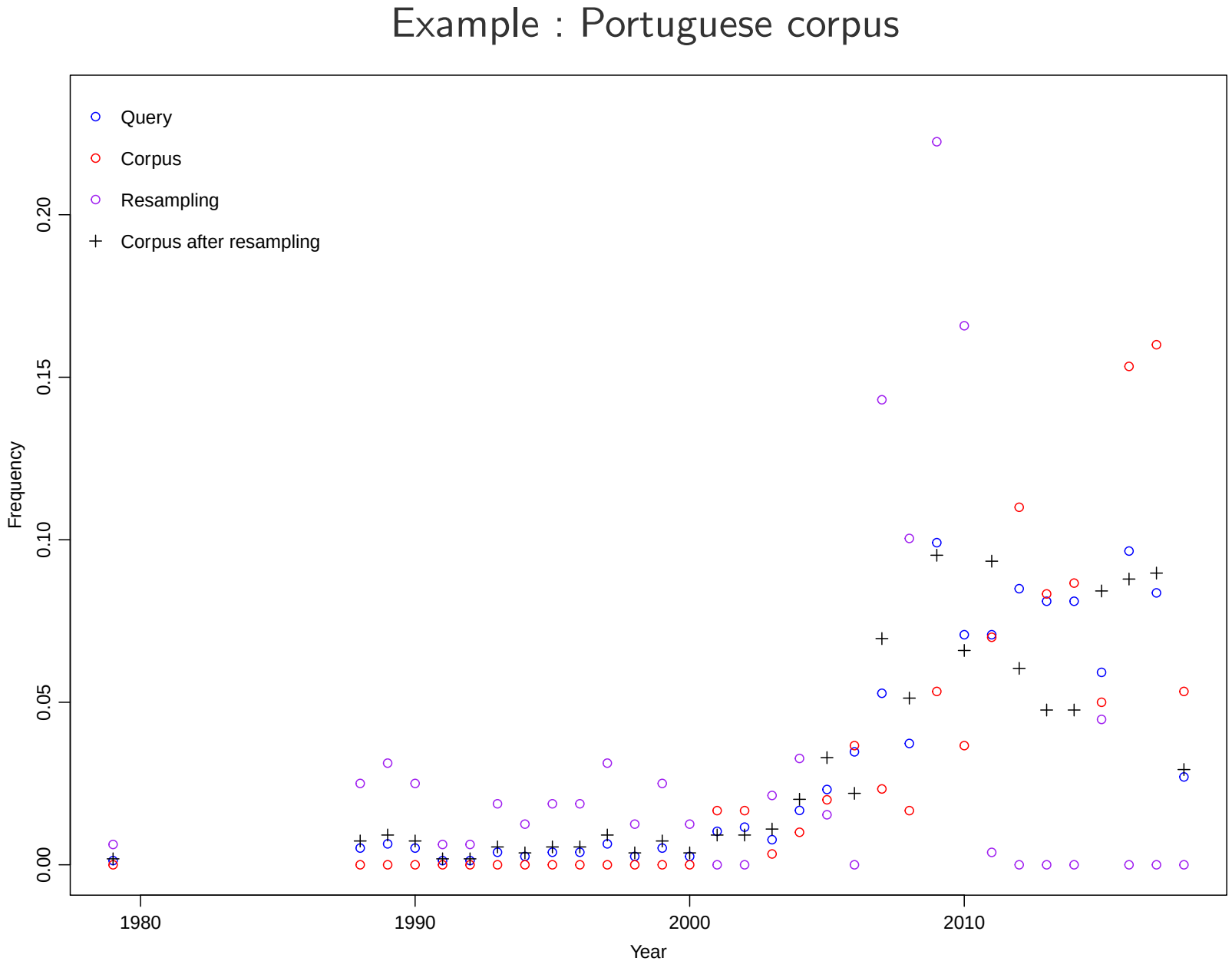
1 Methods

1 Correcting for bias

1. p : probability for paper i to be selected for manual download
2. $p(i \mid \mathrm{Year}) = \Delta[f_{\mathrm{query}}(\mathrm{Year}) - f_{\mathrm{corpus}}(\mathrm{Year})]$
3. $p(i \mid \mathrm{Source}) = \Delta[f_{\mathrm{query}}(\mathrm{Source}) - f_{\mathrm{corpus}}(\mathrm{Source})]$
4. $p(i \mid \mathrm{Year}; \mathrm{Source}) = p(i \mid \mathrm{Year}) \cdot p(i \mid \mathrm{Source})$
5. Systematic sampling with unequal probabilities [Madow, 1949]

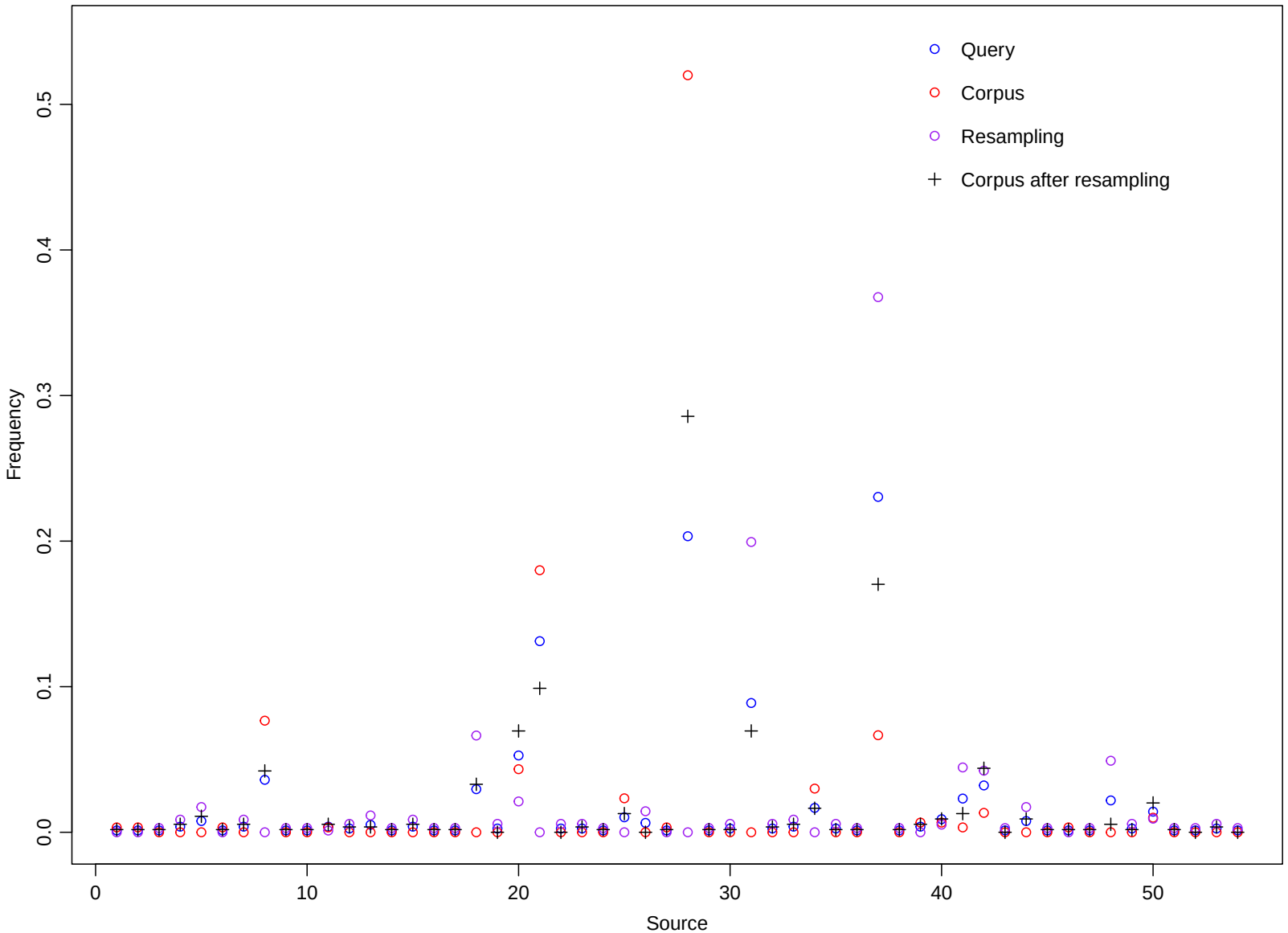
- $\Rightarrow$ Unbiased selection of articles with inclusion probabilities p taking into account evidenced bias between query and corpus

1 Correcting for bias in year



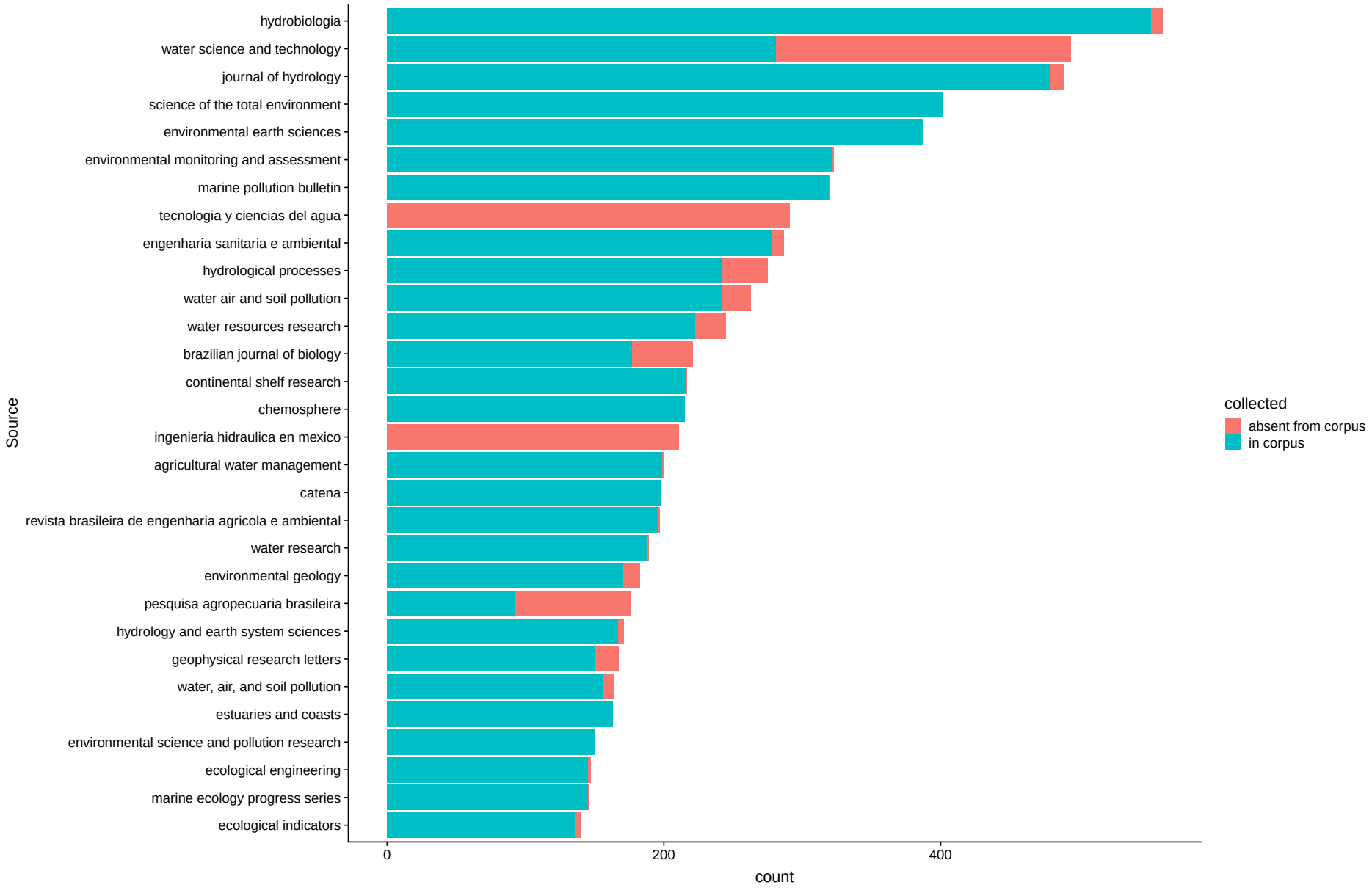
1 Correcting for bias in source

Example : Portuguese corpus



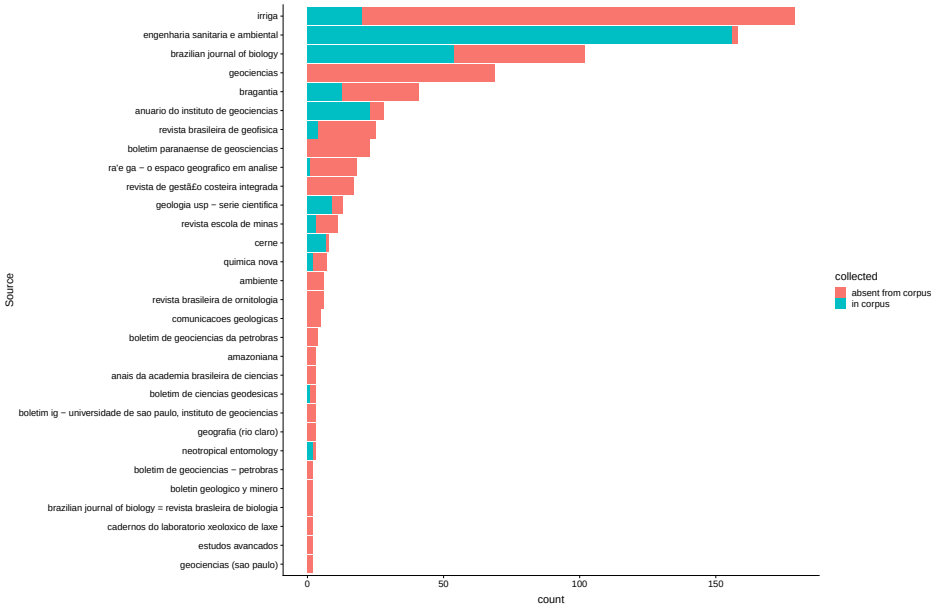
2 Results

2 Baseline : English results - Source

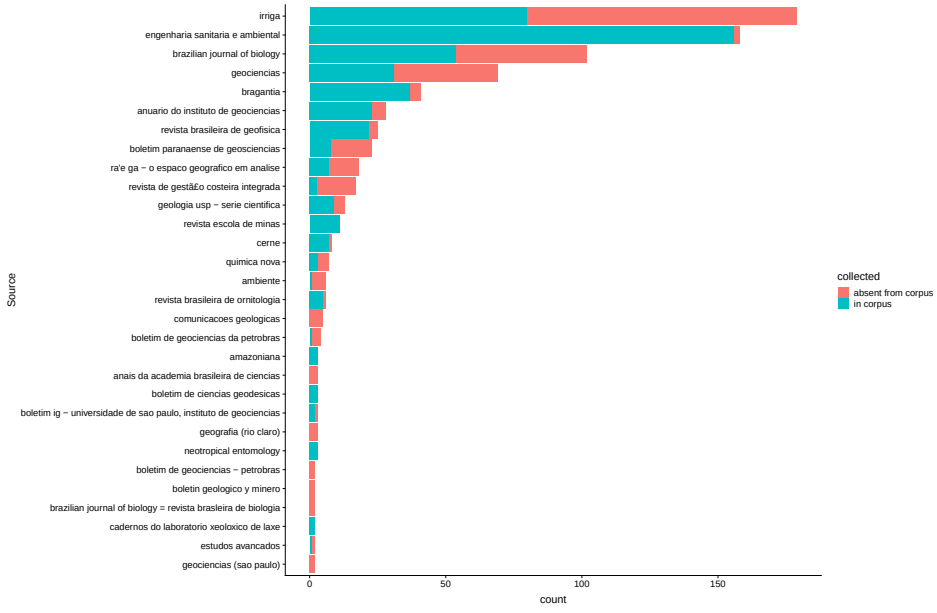


2 Portuguese results - Source

Before download party



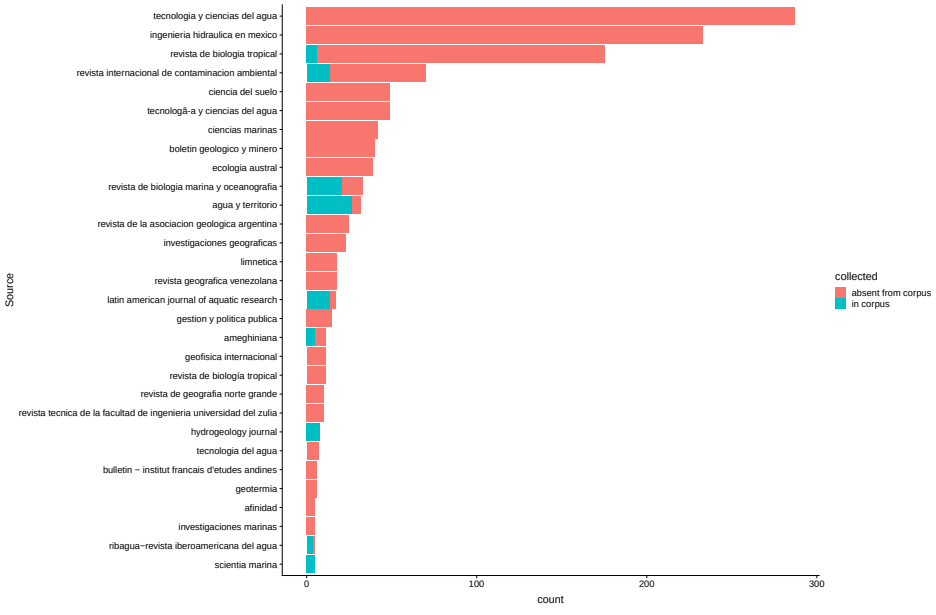
After download party



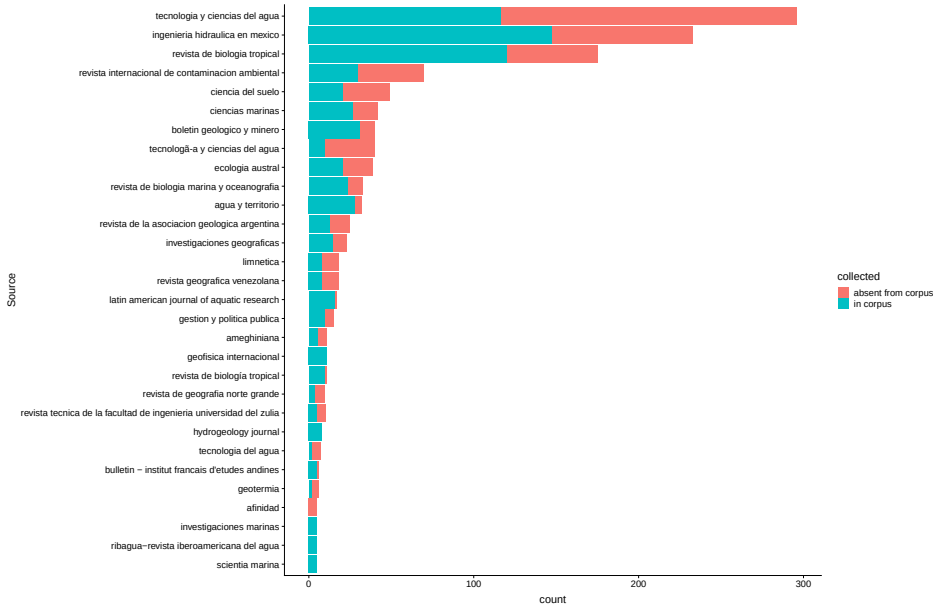
- 77% return rate

2 Spanish results - Source

Before download party

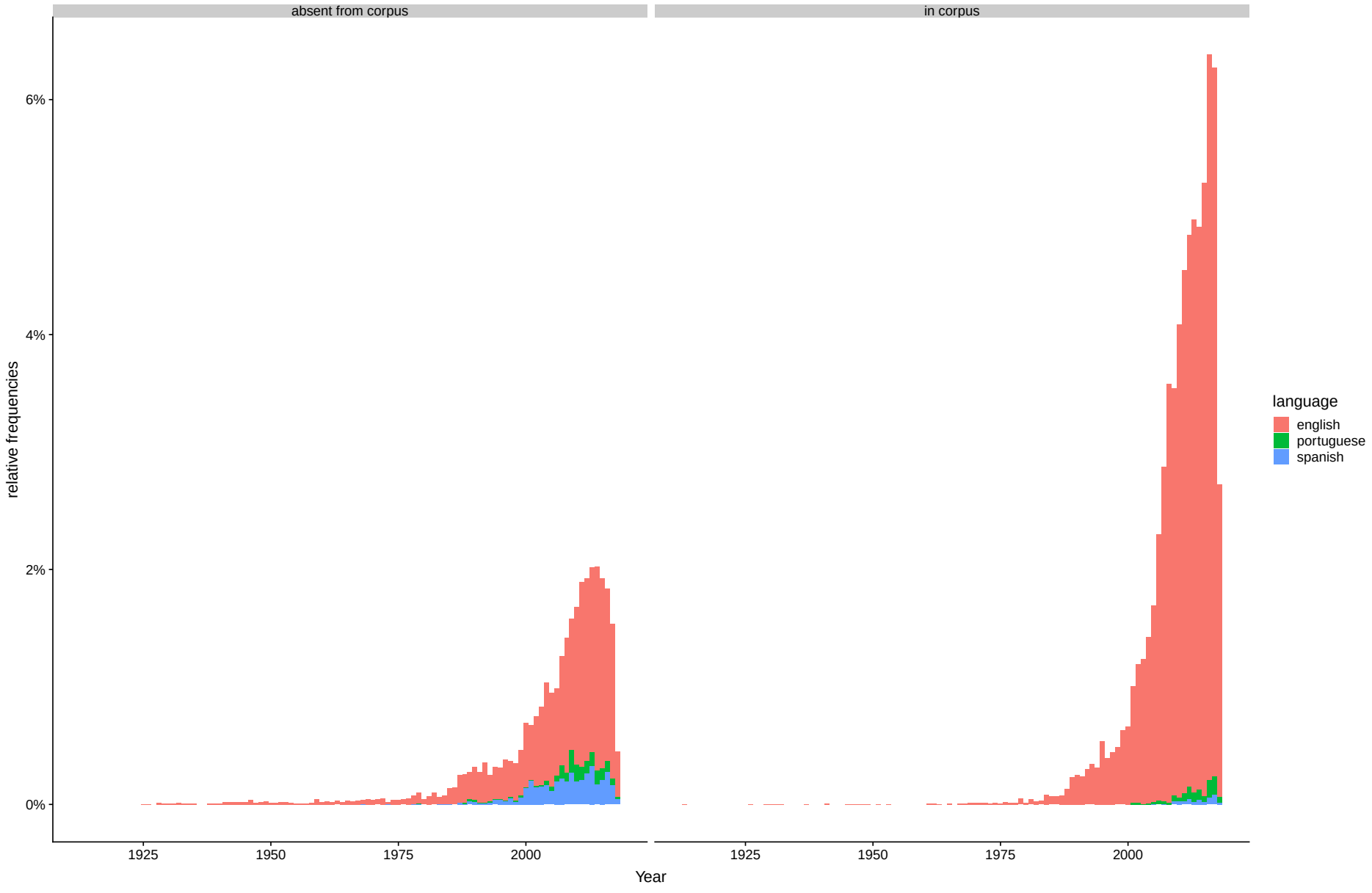


After download party

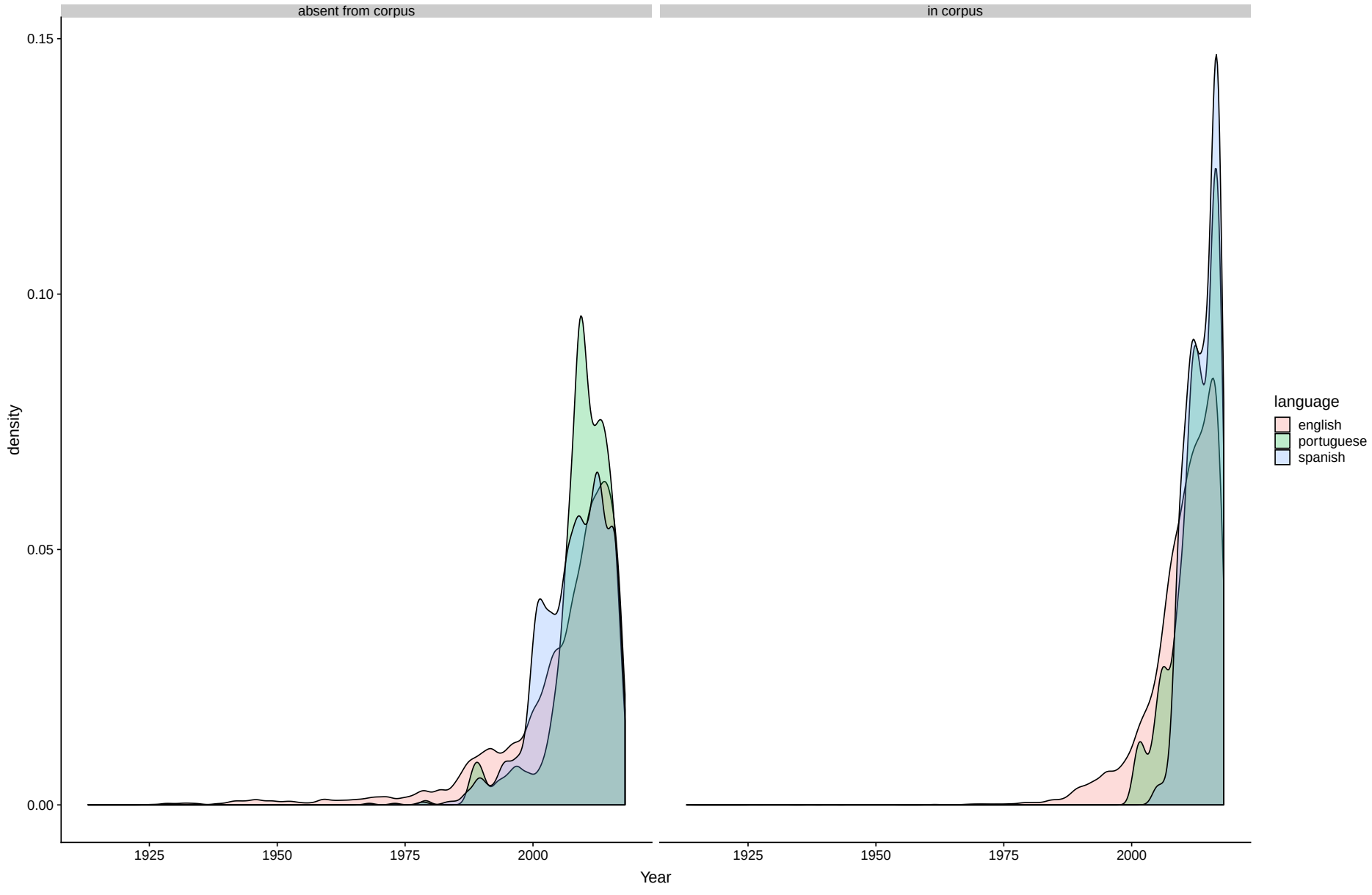


- 80% return rate

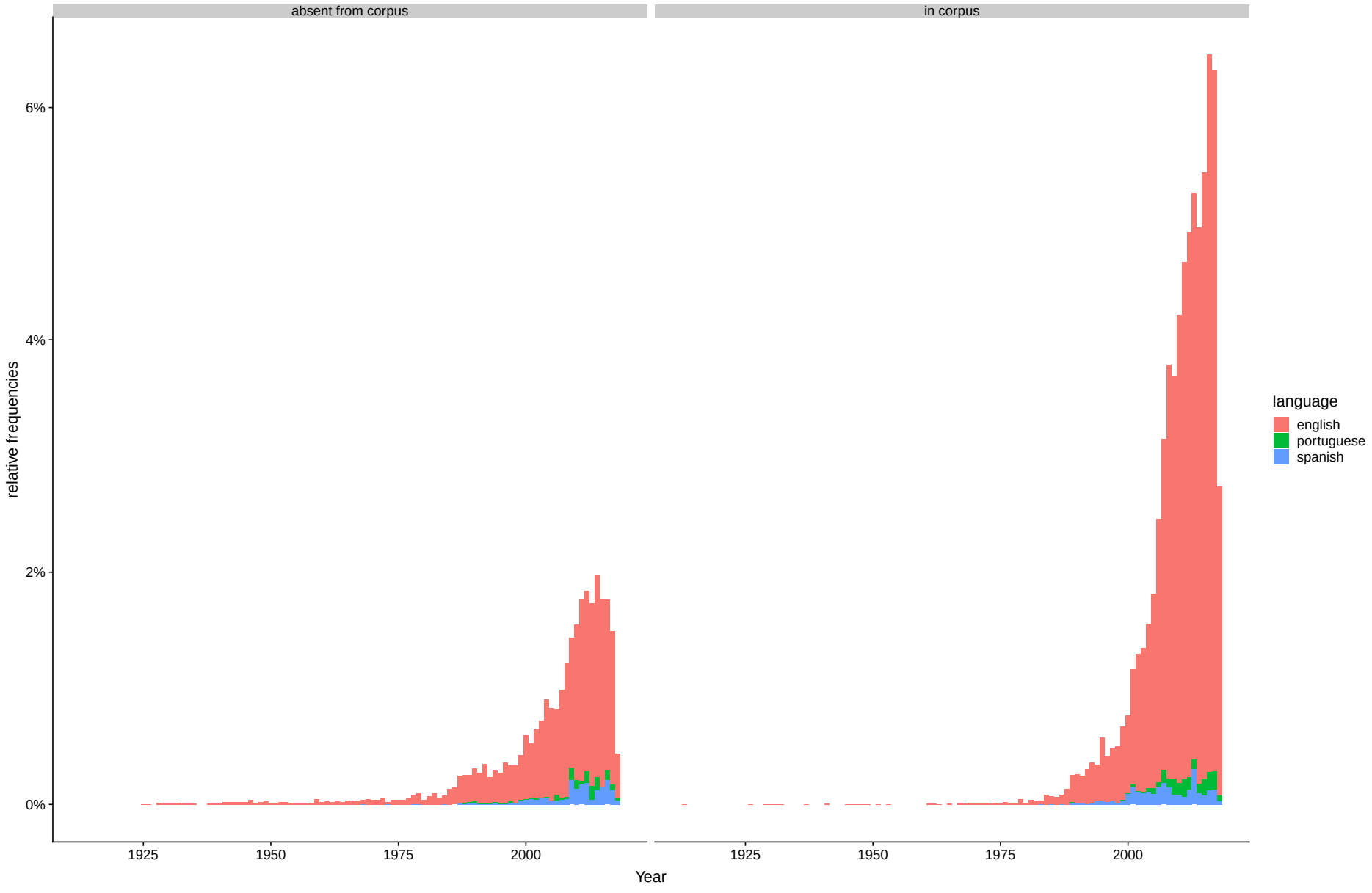
2 Correcting for year : before download party



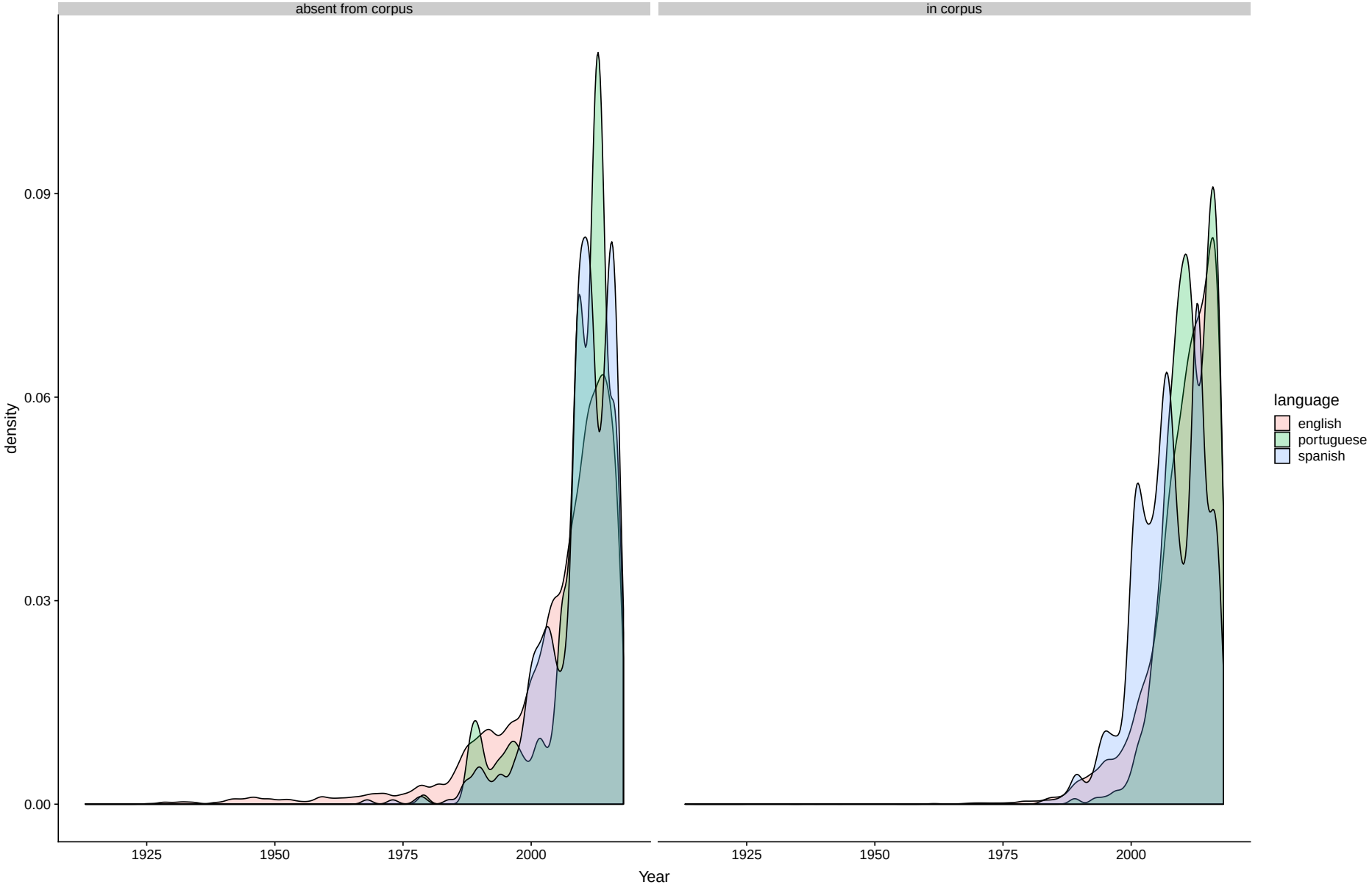
2 Correcting for year : before download party



2 Correcting for year : after download party



2 Correcting for year : before download party



3 Conclusion and next steps

3 Conclusion and next steps

- Bias successfully corrected
- Added
  - 186 Portuguese articles (x 62%) ;
  - 672 Spanish articles (x 550 %)
- Return rate between query and corpus :
  - English : 72%
  - Portuguese : 63%
  - Spanish : 56%
- Number of articles :
  - English : 21 197
  - Portuguese : 486 articles (75 to reach English return rate, x 15%)
  - Spanish : 794 articles (225 to reach English return rate, x 28%)
- Good enough or another download party ?

  (300 papers to download vs. 1 158)