Comparing groups of texts

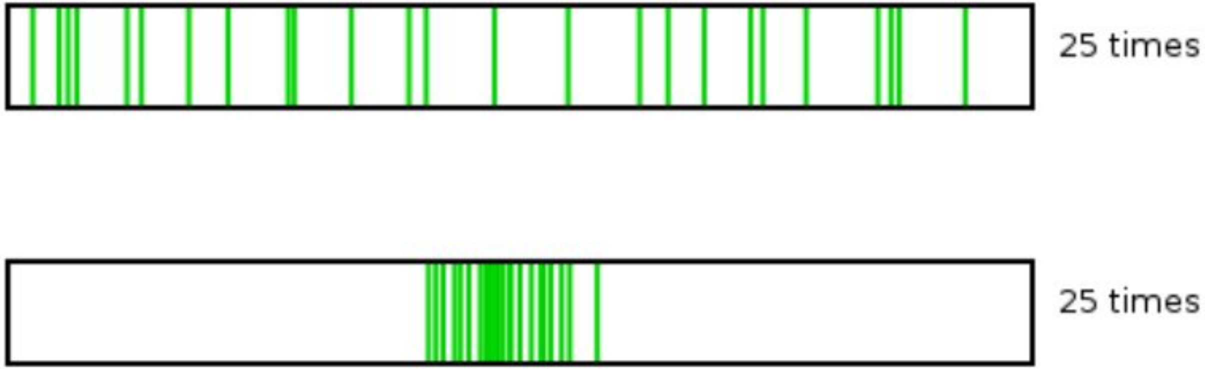
Joanna Byszuk & Maciej Eder
DHSI 2019

About Craig’s Zeta

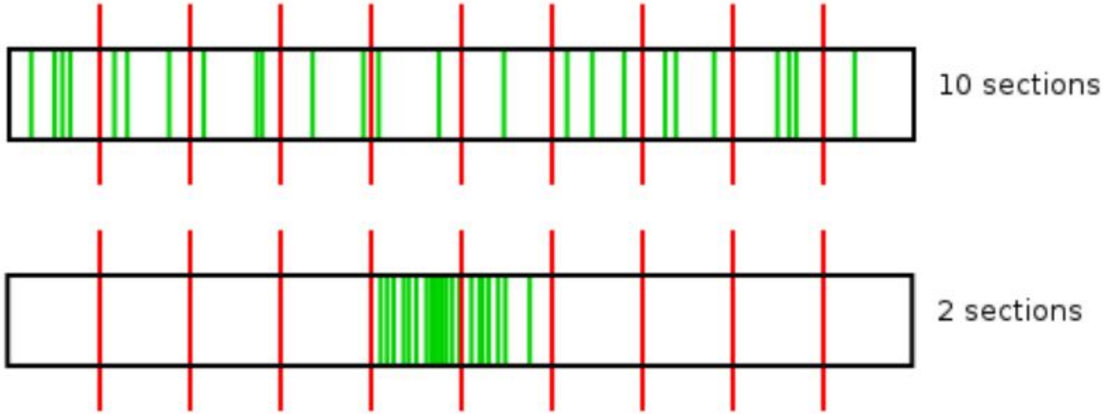
~~Frequencies matter~~

- Zeta focuses on CONSISTENCY of appearance of a word/feature in the text

Distribution matters



Distribution matters



Craig’s Zeta – how to

- Take a corpus, split it in two
- Slice each text into smaller sections
- See if the word is there

Craig’s Zeta

$$\zeta_{(a,b)} = \left(\frac{f_{(a)} - f_{(b)}}{100} \right) + 1 \qquad \zeta_{(a,b)} = \frac{f_{(a)} - f_{(b)}}{f_{(a)} + f_{(b)}}$$

$f(a)$ – frequency in section ‘a’

$f(b)$ – frequency in section ‘b’

Craig's Zeta

in Stylo

Starting the analysis

- Two corpora we hypothesize are different in a significant way

Calling the function

oppose()

Or if it’s not in English:

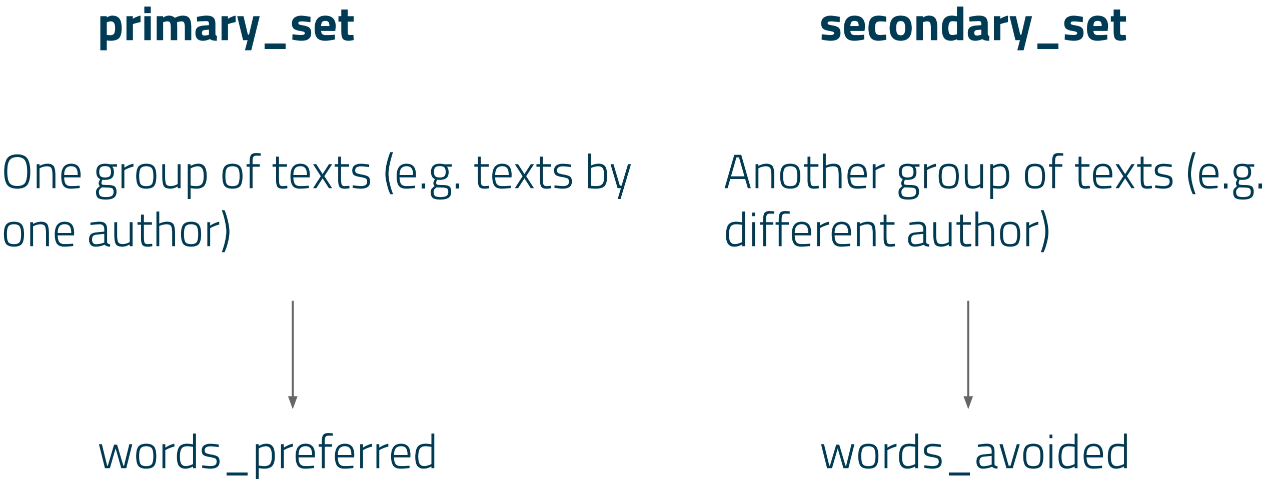
oppose(encoding = "UTF-8", corpus.lang = "Spanish")

Calling the function

Parameters to consider

- Slice length: size (in words) of the samples (5000)
- Slice overlap: (0)
- Method: (Craig's Zeta)
- Visualization: type of graph (Markers / Words)

Comparing corpora in lexis



primary_set

One group of texts (e.g. texts by one author)

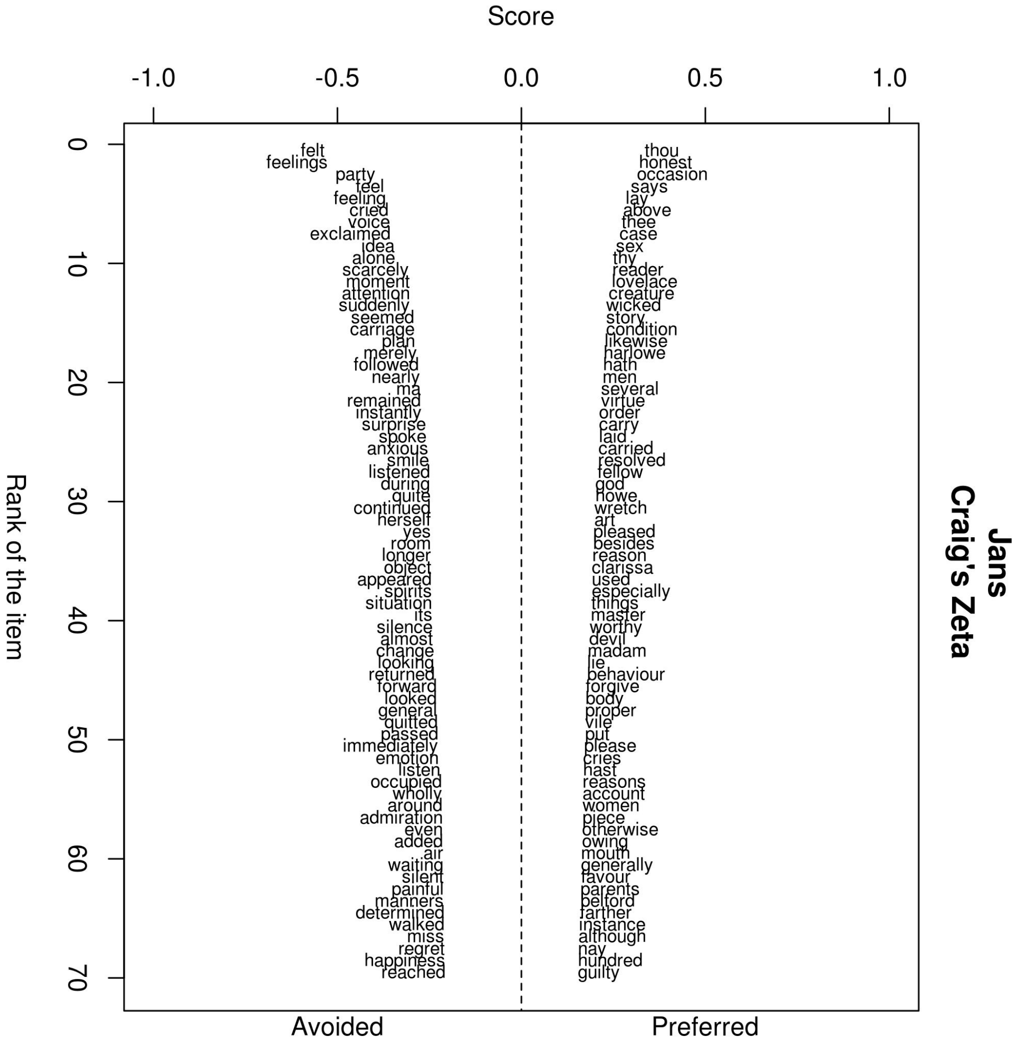
words_preferred

secondary_set

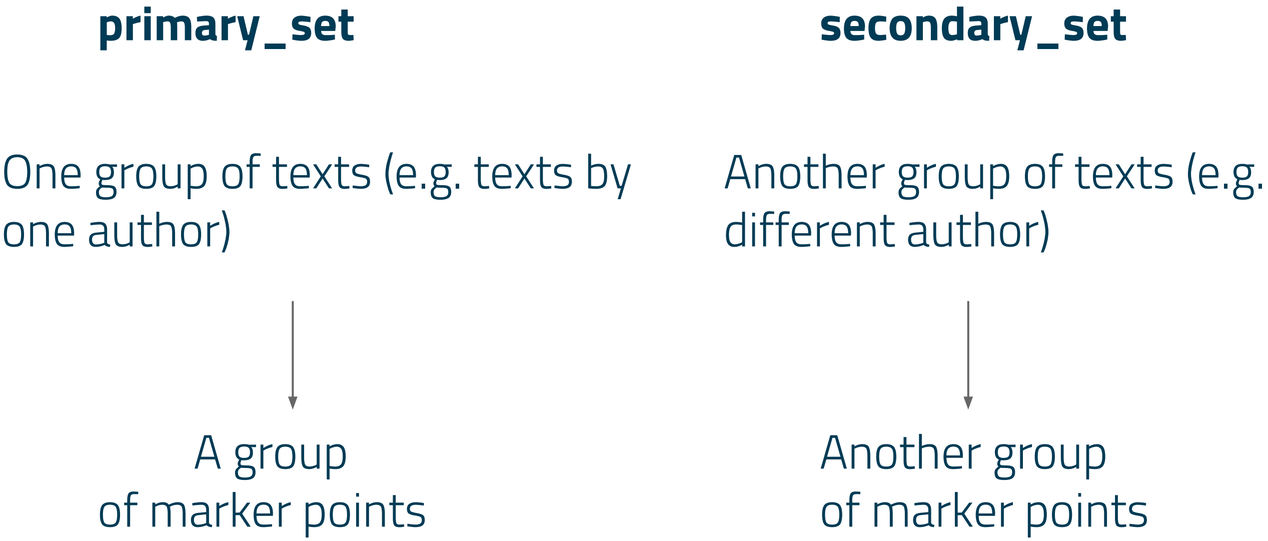
Another group of texts (e.g. different author)

words_avoided

Comparing corpora in lexis



Comparing corpora in markers



primary_set

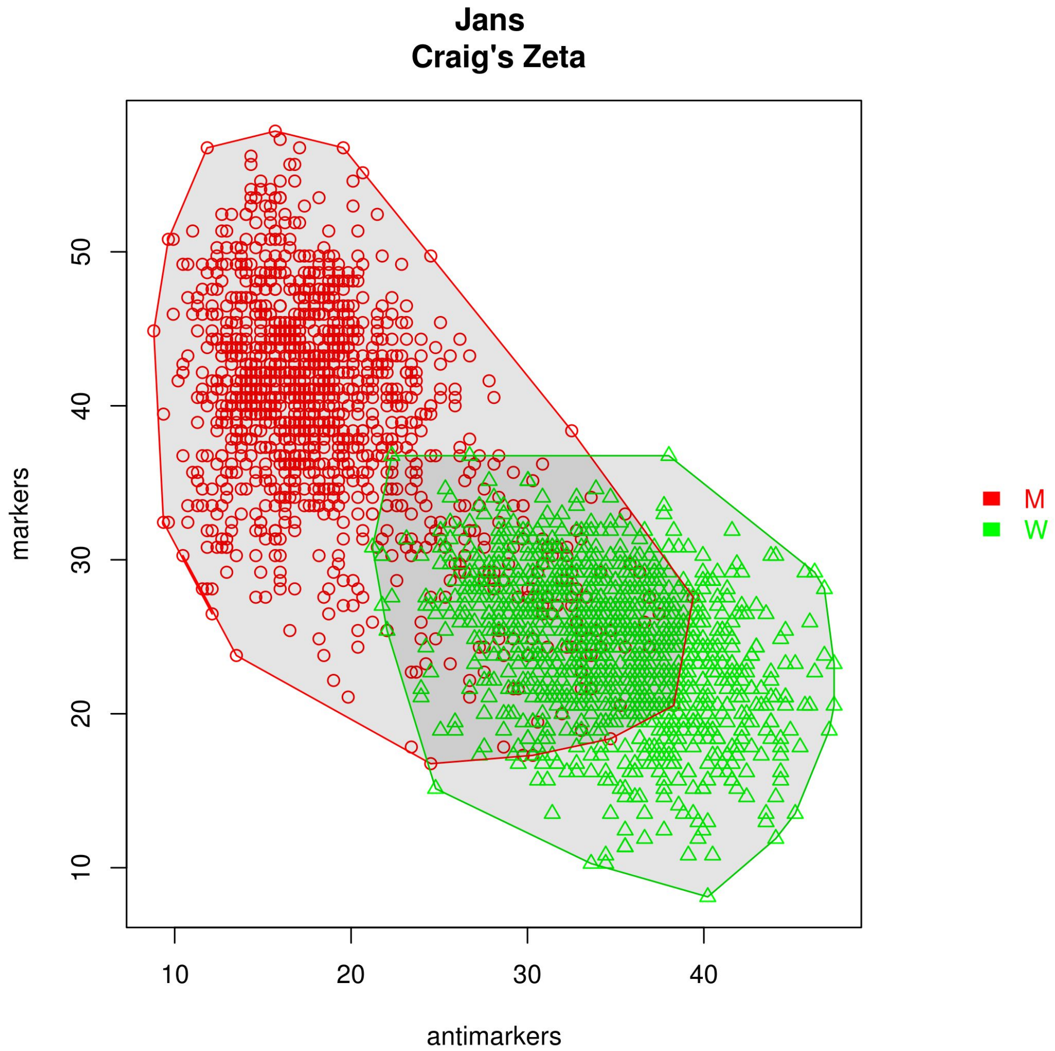
One group of texts (e.g. texts by one author)

A group of marker points

secondary_set

Another group of texts (e.g. different author)

Another group of marker points



Comparing corpora in markers

Applications

Applications

- Comparing discourse markers
- Seeing how distinct two groups *really* are