MAPPING THE RELATIONSHIPS OF CONCEPTS IN TEXT

Simon Musgrave (simon.musgrave@monash.edu)
Brian Zuccala (brian.zuccala@monash.edu)
School of Languages, Literatures, Cultures and Linguistics, Monash University

- Harris (1954:156):
  "difference of meaning correlates with difference of distribution"
- Firth (1957):
  "a word is characterized by the company it keeps"
- Can such distributional relations be made precise?
- The distribution of words can be expressed mathematically as **vectors**
- A vector is a table with a single row
- The vector for any given word records its co-occurrence with other words
  - Each entry in the row corresponds to another word
  - The entry records some information about the co-occurrence of the two words

- Two questions:
  - What is the domain within which co-occurrence is tracked?
  - What information is stored? E.g. is it just the fact of co-occurrence or is it richer information such as distance between words?
- Example with a simple approach:
  - Domain is a sentence
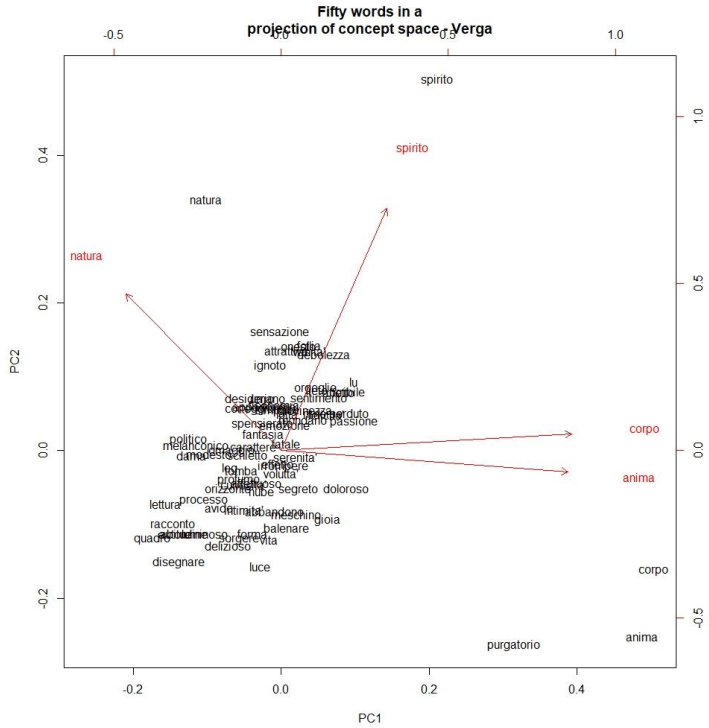 - Information stored is number of times a word occurs

Vector example

	Half	Mushrooms	Onion	The	thinly
a. Thinly SLICE half the onion	1	0	1	1	1
b. SLICE the mushrooms thinly	0	1	0	1	1

- Vectors derived from any large corpus will be very large
- Data is sparse – a very large proportion of the entries are zeroes
- Various algorithms have been developed to reduce the size of the output while preserving information
- One approach reduces the raw vectors to a multidimensional spatial model
- Word2vec uses this approach
  - Word2vec uses neural networks to get from text to spatial model
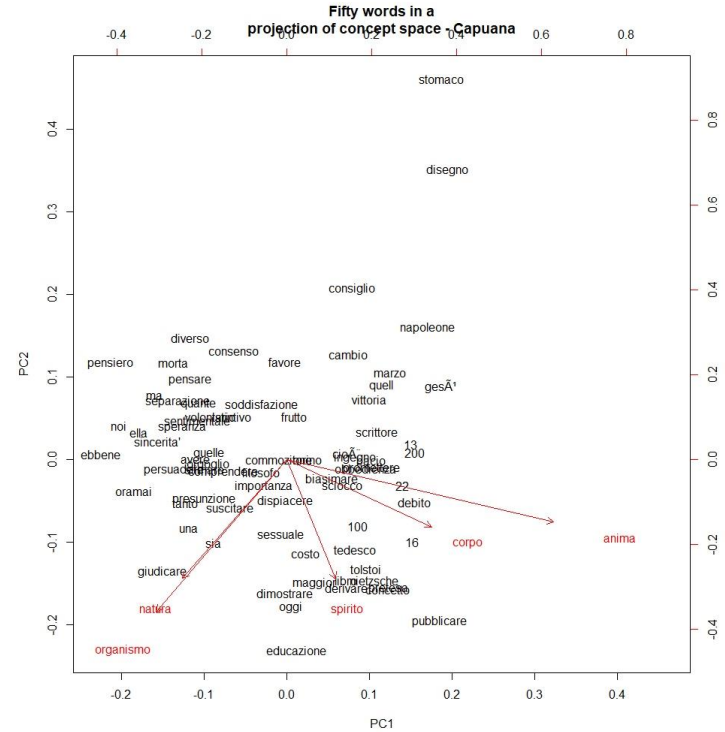- Output is an n-dimensional model which locates all words (lemmas) in relation to each other

- 100 dimensions is still not manageable for visualisation
- We take a small group of words denoting central concepts:
  anima, corpo, spirit, natura, organism
- 50 words which are all close to these concepts are extracted
- PCA gives us the two components in that groups of data which:
  - Are orthogonal to each other
  - Account for the greatest amount of the variance in the data
- This can be plotted in 2D

- We could try to attribute meaning to the principal components, but it is not useful
  - Cf. topic models – the technique tells us words that are associated, but we have to try to give meanings to those groups
- Even if we can attribute meaning to the PCs in one visualisation, we know that they will not be the same in comparing visualisations
- Therefore we cannot compare directionality across visualisations
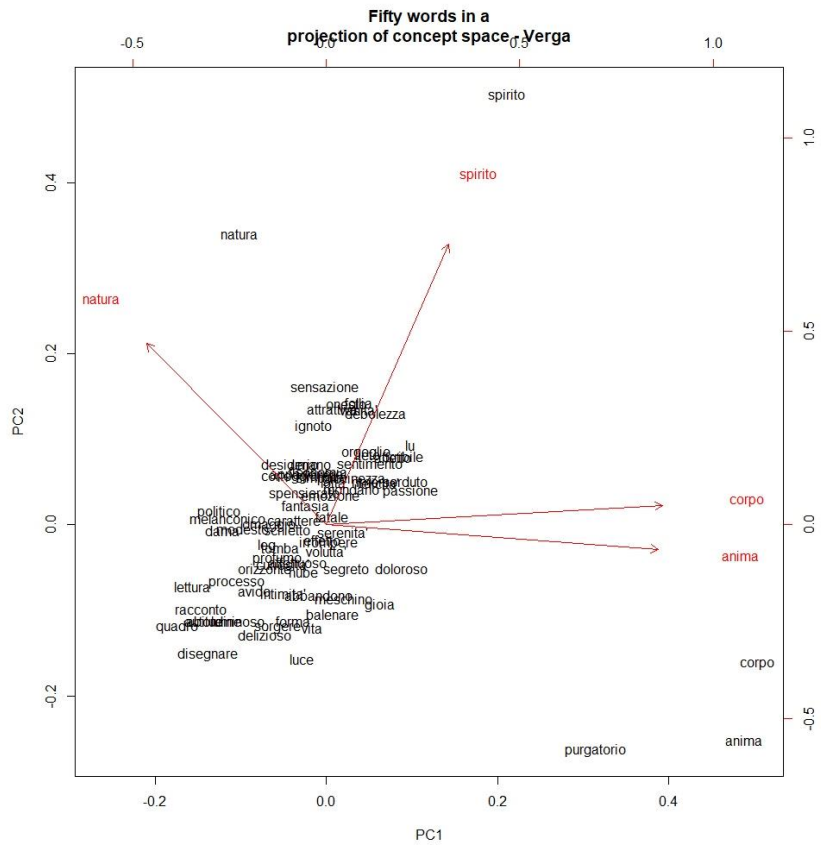- Clustering and relative configurations are what we can try to interpret
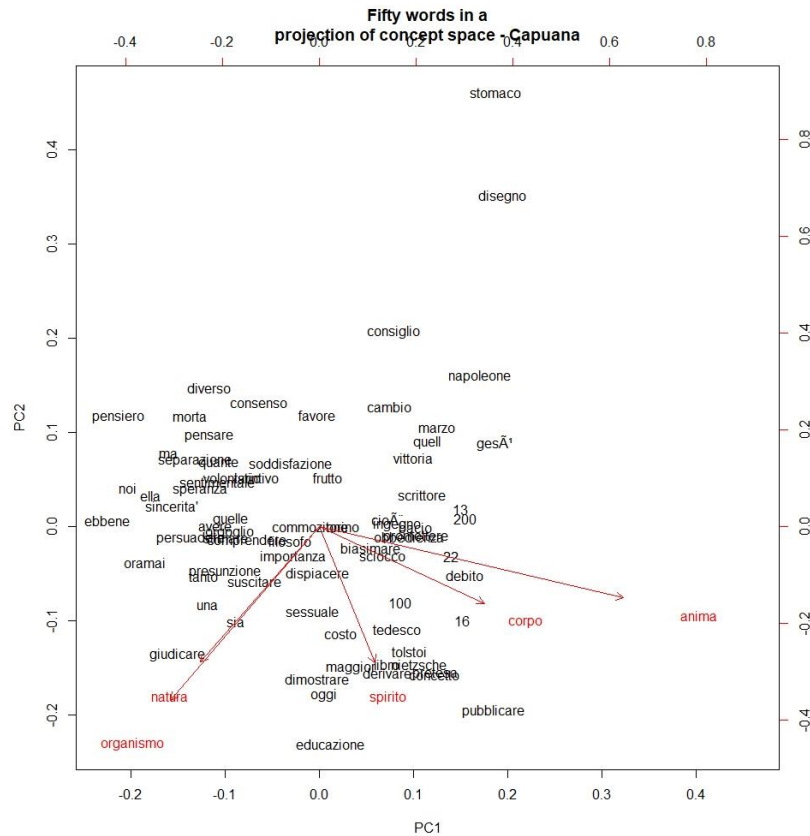
Visualising 'concept space'

Verga

Capuana

Verga

Fifty words in a
projection of concept space - Verga

Capuana

Fifty words in a projection of concept space - Capuana

Comparison – 1: Density

- Verga word cloud is denser, Capuana more dispersed
- Positivism peaks in Italian culture around Verga's time
- Seen in cultural studies as:
  - a deterioration of the epistemological unity of the world
  - a subsequent fragmentation of the psychological unity of the human subject
- This is reflected in our visualisation of concept space

Visualising 'concept space'
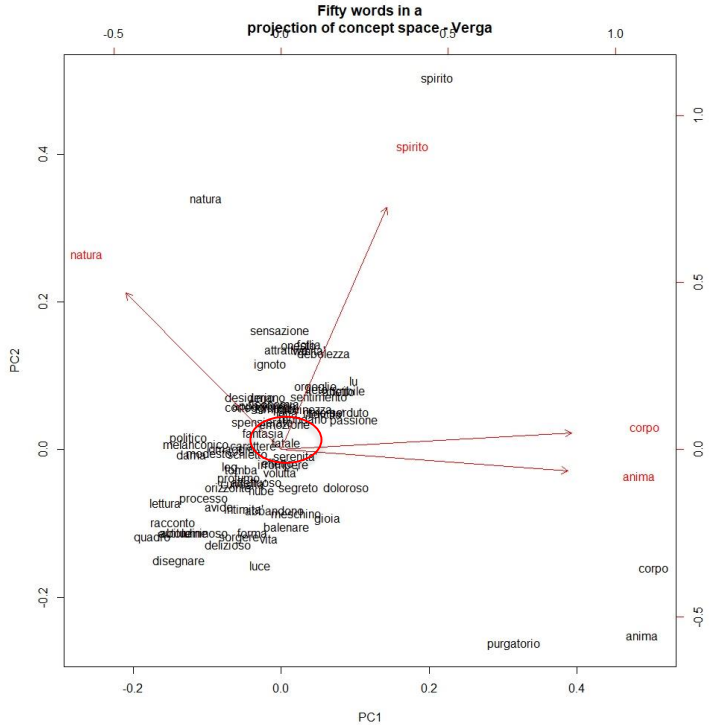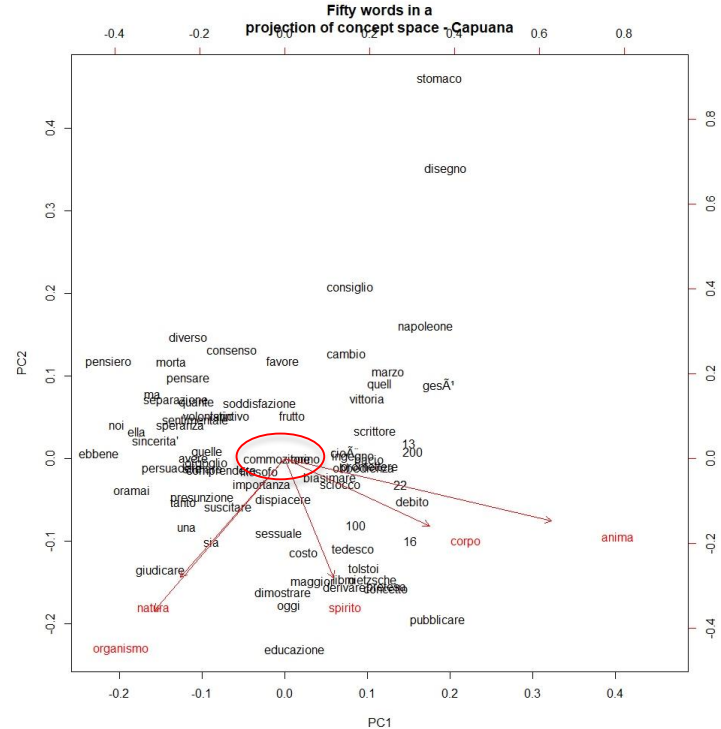
Verga

Capuana

- Verga's concept space centres on *fatale*
  - (Many of) Verga's characters are subaltern, trapped by inescapable socio-economic forces
- Capuana's concept space centres on *commozione*
  - For Capuana one of the crucial aims of literature itself
  - Art must go about the representation of its subjects in such a way as to reach the reader's nervous system and trigger their emotions

13

- Students often have problems keeping track of the '-isms' of literature

- These visualisations can assist in showing:
  - Differences between authors
  - Differences across time