NLP

Course of DSSC Master degree - University of Trieste

Matilde Trevisani

DEAMS

2020/05/25 (updated: 2020-06-12)

Content Statistical analysis

Content Statistical analysis

1. Content mapping
    - Correspondence analysis
2. Topic detection
    - LDA
    - Reinert Method

Content Mapping

Is it possible placing all the texts (or subcorpora) on a map based on their lexical similarity?

And is it possible to find a graphical representation that uncovers similarities and differences between texts, between words and between texts and words?

Correspondence analysis (CA) is a method that allows you to transform lexical profiles into positions on a Cartesian plane.

- CA, operating on rows (words) and columns (texts or subcorpora) of a word $\times$ text matrix (contingency table), transforms the frequencies (cells of the matrix) in coordinates on a multidimensional system of Cartesian axes.
- CA transforms an appropriate distance - χ^2 distance - calculated between pairs of rows and pairs of columns - into a (weighted) Euclidean distance, whence projects texts and/or words on a Cartesian plane (dimensional reduction).
- The procedure is based on a Multidimensional scaling on the χ^2-based distance matrices or, equivalently, on a singular value decomposition.

Lexical profile

forme	occorrenze	autore01	autore02	autore03	...	autore39	autore40
di	312.712	2.988	8.602	3.117	...	3.617	8.100
e	280.248	1.370	12.003	4.495	...	2.989	6.924
che	244.772	1.559	6.746	3.428	...	2.050	9.194
la	198.228	1.745	7.220	3.288	...	2.526	6.029
a	186.667	1.680	6.480	2.271	...	2.421	5.179
il	171.147	1.778	5.922	2.585	...	2.293	4.531
non	158.221	899	5.329	1.447	...	1.901	5.013
un	150.893	1.059	5.795	2.082	...	1.888	4.493
in	127.484	1.457	3.560	2.133	...	1.534	3.573
...	...	...	...	...	...	...	...
cosa	20.409	85	711	157	...	121	744
tutto	19.992	112	662	297	...	89	464
senza	18.895	165	488	211	...	244	526
...	...	...	...	...	...	...	...
casa	15.667	80	594	430	...	149	332
...	...	...	...	...	...	...	...

Lexical profile

forme →	di	e	che	la	a	il	non	un	in	...	cosa	tutto	senza	..	casa	...
autore01	2.988	1.37	1.559	1.745	1.68	1.778	899	1.059	1.457	...	85	112	165	..	80	...
autore02	8.602	12.003	6.746	7.22	6.48	5.922	5.329	5.795	3.56	...	711	662	488	..	594	...
autore03	3.117	4.495	3.428	3.288	2.271	2.585	1.447	2.082	2.133	...	157	297	211	..	430	...
...	...	...	...	...	...	...	...	...	...	...	...	...	...	..	...	...
autore39	3.617	2.989	2.05	2.526	2.421	2.293	1.901	1.888	1.534	...	121	89	244	..	149	...
autore40	8.1	6.924	9.194	6.029	5.179	4.531	5.013	4.493	3.573	...	744	464	526	..	332	...
occ.es	312.712	280.248	244.772	198.228	186.667	171.147	158.221	150.893	127.484	...	20.409	19.992	18.895	..	15.667	...

Lexical profile

	1,000 2-grams				1,000 3-grams				1,000 2-word-grams				1,000 most freq. words			
	di	*ch*		*pe*	*del*	*che*		*per*	*di cui*	*con gioia*		*setto nasale*	*casa*	*cosa*		*madre*
	cbg0001	cbg0002	...	cbg1000	ctg0001	ctg0002	...	ctg1000	wbg0001	wbg0002	...	wbg1000	mfw0001	mfw0002	...	mfw1000
chunk0001	1.57	1.57	...	0.79	0.39	0.39	...	0.39	0.00	0.00	...	0.00	4.76	1.90	...	0.00
chunk0002	2.27	1.23	...	0.09	0.47	0.57	...	0.28	0.88	0.00	...	0.00	1.94	1.94	...	0.00
chunk0003	0.11	2.44	...	1.38	0.32	0.43	...	0.21	0.89	0.44	...	0.00	3.88	4.85	...	0.00
chunk0004	1.28	0.79	...	1.67	0.59	0.49	...	0.29	0.43	0.00	...	0.00	2.93	1.95	...	0.49
chunk0005	1.42	2.25	...	1.13	0.51	0.61	...	0.21	0.00	0.44	...	0.00	4.00	1.96	...	0.49
chunk0006	1.79	1.59	...	0.10	0.40	0.10	...	0.10	2.21	0.00	...	0.00	1.94	6.31	...	0.00
chunk0007	2.37	0.09	...	0.85	0.47	0.28	...	0.66	1.29	0.00	...	0.00	4.74	3.79	...	0.47
....	1.48	2.31	...	1.39	0.74	0.28	...	0.28	0.42	0.84	...	0.00	4.37	0.49	...	0.00
....	1.61	1.01	...	1.41	0.71	0.40	...	0.40	0.43	0.43	...	0.00	3.43	2.94	...	0.00
....	1.04	2.08	...	1.23	0.57	0.09	...	0.38	0.44	0.44	...	0.88	2.45	2.45	...	0.00
....	1.48	1.21	...	0.74	0.84	0.56	...	0.37	0.42	0.42	...	0.00	5.80	1.93	...	0.00
....	1.75	0.10	...	1.46	0.29	0.29	...	0.29	0.00	0.00	...	0.46	6.40	1.97	...	0.00
....	1.27	1.45	...	0.09	0.18	0.55	...	0.18	0.90	0.00	...	0.45	4.88	2.93	...	0.00
....	0.74	1.58	...	0.95	0.21	0.53	...	0.11	0.46	0.00	...	0.00	4.46	3.47	...	0.00
....	1.37	1.96	...	1.47	0.39	0.29	...	0.88	0.00	0.00	...	0.45	8.00	2.50	...	0.00
....	1.22	0.10	...	1.52	0.51	0.20	...	0.51	0.00	0.00	...	0.00	3.00	0.50	...	0.00
chunk9512	0.10	1.73	...	0.96	0.19	0.10	...	0.67	0.44	0.88	...	0.44	4.00	2.50	...	1.00
chunk9513	1.16	1.65	...	1.16	0.58	0.39	...	1.07	0.44	0.44	...	0.00	4.78	0.96	...	0.00
chunk9514	1.45	1.77	...	2.28	0.21	0.31	...	0.21	0.46	0.00	...	0.00	4.93	1.97	...	0.00

Correspondence analysis

Correspondence analysis (**CA**) is a method for explorative data analysis (EDA) useful for studying the joint distribution of two (or more) categorical variables.

It represents the association structure existing between two (or more) categorical variables through a simple graphs that place (project) the variable categories on Cartesian planes.

In text analysis perspective under a bag-of-words approach, CA can be used to map the contents of a corpus: if the contingency table is made up of a matrix of word-by-text frequencies, the system of relationships existing

- between words,
- between texts and
- between texts and words

can be represented on Cartesian planes.

Of course, CA cannot represent all the corpus characteristics but it is able to uncover the predominant latent dimensions.

Example

Suppose that a corpus consists of all the articles published in the annual volumes of one scientific journal and that the articles are organized in subcorpora by year.

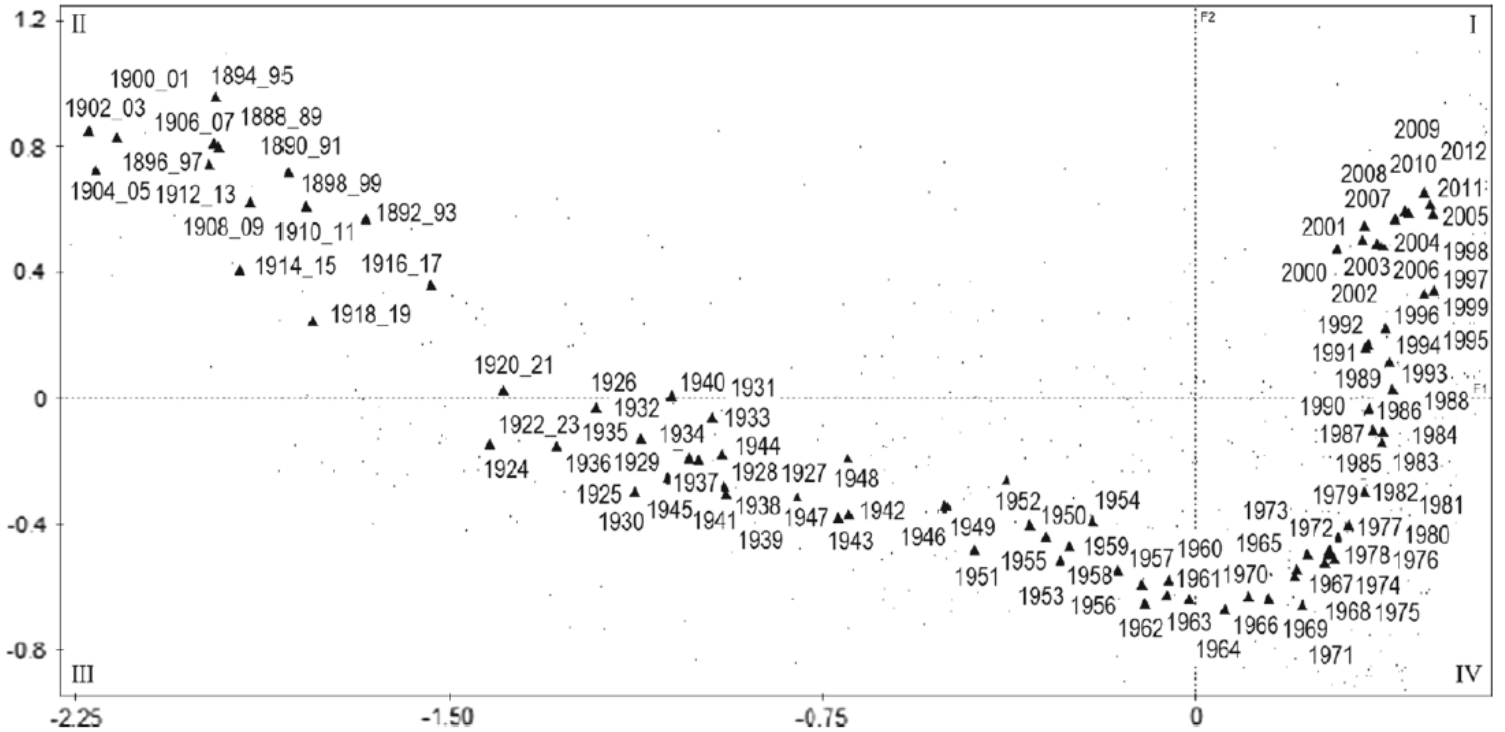
We also suppose that the keywords of the articles have been selected from the corpus vocabulary.

The contingency table represents the frequencies (cells) of the keywords (rows) in the different years (columns).

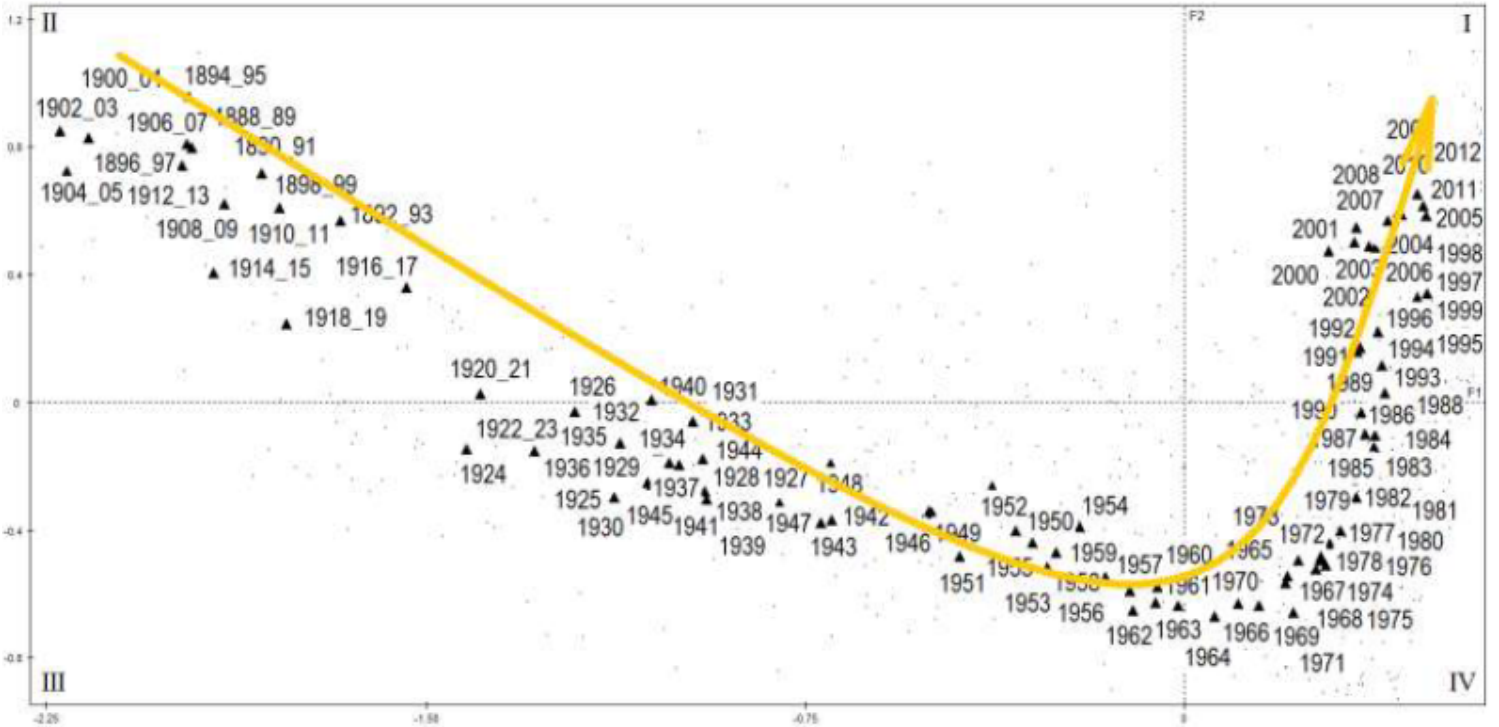
CA represents the system of association between keywords (rows), between years (columns) and between words and years.

Example (continued)

In the case of the corpus made of journal articles and the keyword per year matrix, CA can verify if there is an apparent temporal pattern.



Example (continued)



CA

In the simplest version, CA works on a lexical contingency table where rows are related to a selection of words (m word-types $w_1, \ldots, w_m$) and columns are related to texts or subcorpora (p texts $t_1, \ldots, t_p$).

Each cell reports the number of occurrences n_{ij} of the i-th word (row) in the j -th text (column).

Example of word $\times$ text - lexical contingency table (table of joint frequencies)

	t_1	t_2	$\cdots$	t_j	$\cdots$	t_p	
w_1	n_{11}	n_{12}	$\cdots$	n_{1j}	$\cdots$	n_{1p}	$n_{1\cdot}$
w_2	n_{21}	n_{22}	$\cdots$	n_{2j}	$\cdots$	n_{2p}	$n_{2\cdot}$
$\vdots$	$\vdots$	$\vdots$		$\vdots$		$\vdots$	
w_i	n_{i1}	n_{i2}	$\cdots$	n_{ij}	$\cdots$	n_{ip}	$n_{i\cdot}$
$\vdots$	$\vdots$	$\vdots$		$\vdots$		$\vdots$	
w_k	n_{k1}	n_{k2}	$\cdots$	n_{kj}	$\cdots$	n_{kp}	$n_{k\cdot}$
$\vdots$	$\vdots$	$\vdots$		$\vdots$		$\vdots$	
w_m	n_{m1}	n_{m2}	$\cdots$	n_{mj}	$\cdots$	n_{mp}	$n_{m\cdot}$
	$n_{\cdot 1}$	$n_{\cdot 2}$	$\cdots$	$n_{\cdot j}$	$\cdots$	$n_{\cdot p}$	n

Examining dependence in a 2-way table

In elementary statistics courses we learn that it is of interest evaluate whether two characters (categorical variables) are independent.

Independence implies that conditional distributions are equal, whence equal to marginal distribution.

Hence, detecting differences between conditional distributions leads to infer the deviation of the observed data from the theoretic hypothesis of independence.

Test of Independence (based on the X^2 Pearson statistics) is ordinarily used.

Row and column profiles (points)

Conditional distributions by row (word):

$X \,/\, Y$	y_1	$\cdots$	y_j	$\cdots$	y_J	massa
x_1	$\frac{n_{11}}{n_{1.}}$	$\vdots$	$\frac{n_{1j}}{n_{1.}}$	$\vdots$	$\frac{n_{1J}}{n_{1.}}$	$\frac{n_{1.}}{n} = f_{1.}$
$\cdots$	$\cdots$	$\vdots$		$\vdots$	$\cdots$	$\cdots$
x_i	$\frac{n_{i1}}{n_{i.}}$	$\vdots$	$\frac{n_{ij}}{n_{i.}}$	$\vdots$	$\frac{n_{iJ}}{n_{i.}}$	$\frac{n_{i.}}{n} = f_{i.}$
$\cdots$	$\cdots$	$\vdots$		$\vdots$	$\cdots$	$\cdots$
x_I	$\frac{n_{I1}}{n_{I.}}$	$\vdots$	$\frac{n_{Ij}}{n_{I.}}$	$\vdots$	$\frac{n_{IJ}}{n_{I.}}$	$\frac{n_{I.}}{n} = f_{I.}$

mass - given by the marginal relative frequencies - expresses the relative importance of rows (columns).

Row and column profiles (points)

Conditional distributions by column (text):

$X \,/\, Y$	y_1	$\cdots$	y_j	$\cdots$	y_J	
x_1	$\frac{n_{11}}{n_{\cdot 1}}$	$\vdots$	$\frac{n_{1j}}{n_{\cdot j}}$	$\vdots$	$\frac{n_{1J}}{n_{\cdot J}}$	
$\cdots$	$\cdots$	$\vdots$	$\cdots$	$\vdots$	$\cdots$	
x_i	$\frac{n_{i1}}{n_{\cdot 1}}$	$\vdots$	$\frac{n_{ij}}{n_{\cdot j}}$	$\vdots$	$\frac{n_{iJ}}{n_{\cdot J}}$	
$\cdots$	$\cdots$	$\vdots$	$\cdots$	$\vdots$	$\cdots$	
x_I	$\frac{n_{I1}}{n_{\cdot 1}}$	$\vdots$	$\frac{n_{Ij}}{n_{\cdot j}}$	$\vdots$	$\frac{n_{IJ}}{n_{\cdot J}}$	
massa	$\frac{n_{\cdot 1}}{n} = f_{\cdot 1}$		$\frac{n_{\cdot j}}{n} = f_{\cdot j}$		$\frac{n_{\cdot J}}{n} = f_{\cdot J}$	

Rows and colums of such matrices are called **row profiles** and **column profiles** (or, more simply, **row points** and **column points**).

Similarity (dissimilarity) betwen row profiles (column profiles) indicates independence (association).

Vector interpretation

Rows and columns of the contingency table can also be considered as vectors, that is, as **points** in multidimensional space.

Each word is a p-dimensional vector in the text space.

Each text is an m-dimensional vector in the word space.

- X_r: matrix of row-points in a space of p dimensions;
  - row-points with relative importance r (mass)
- X_c: matrix of column-points in a space of m dimensions
  - column-points with relative importance c (mass)

Average profiles (by weighted mean of row/column profiles with weight given by row/column mass)

$$c = X_r^T r \qquad r = X_c^T c$$

that are the marginal profiles c and r are the **centroids** of these spaces.

Geometric interpretation

This geometric representation of profiles suggests to explore if there is a low-dimensional system of coordinates that preserve (with the least loss of information) the structure of the distances between them.

The distance between two vectors is measured through a **weighted Euclidean distance** (Balbi and Misuraca, 2005; Murtagh, 2005) which compares the lexical profiles taking into account the size of the texts (text mass) and the total occurrences of the words (word mass).

Dimensionality reduction techniques

This approach is the same that underlies the **multidimensional scaling** analysis as well as the **biplot** (referring to space of row points that preserve the distance structure).

Note

But it is also the same approach underlying the **principal component analysis** if we replace the concept of distance with that of covariance / correlation.

Distance between two words (rows)

The purpose of CA is to translate similarity between categories (words and texts) in a graph where the most similar categories are placed in close positions on the space delimited by the Cartesian axes.

If we consider words, it is quite intuitive to think that the similarity between two words depends on "how similar the occurrences are" in the two rows of the table, that is, how similar they are in terms of presence, absence and occurrence in the texts of the corpus.

If two words tend to be used in the same texts and with similar relative frequency, they have a similar profile.

Two words with identical profile will have zero distance, that is, they will be represented on a graph as two overlapping points.

χ^2 distance

The intuitive concept of (dis)similarity between profiles of two words w_i and w_l is translated into a distance ("chi-square" χ^2 distance) calculated for each pair of words as follows:

$$d^2_{il(r)} = \sum_{j=1}^{p} \frac{1}{f_{\bullet j}} \left(\frac{n_{ij}}{n_{i\bullet}} - \frac{n_{lj}}{n_{l\bullet}} \right)^2$$

where $f_{\bullet j} = \frac{n_{\bullet j}}{n}$ is the **mass** of column j (given by the marginal relative frequency of text j).

A weighted distance

Why do we use the reciprocal of the mass?

Since each component of the summation is influenced by the marginal relative frequency $\frac{n_{\bullet j}}{n} = f_{\bullet j}$, it seems appropriate to relativize each component with such frequency.

The use of the reciprocal of the mass is ***variance - standardizing***: it compensates for the larger variance for high frequencies and smaller variance for low frequencies.

If this standardization were not applied, differences between larger proportions would tend to be larger and thus to dominate in distance computation, while differences between smaller proportions would tend to be overwhelmed.

Distance between two texts (columns)

The reasoning can be repeated taking into account the (dis) similarity between pairs of texts and considering the profiles of two columns.

Two generic texts t_j and t_l are similar if they have a similar lexical profile, i.e., if they include the same words with a similar relative frequency.

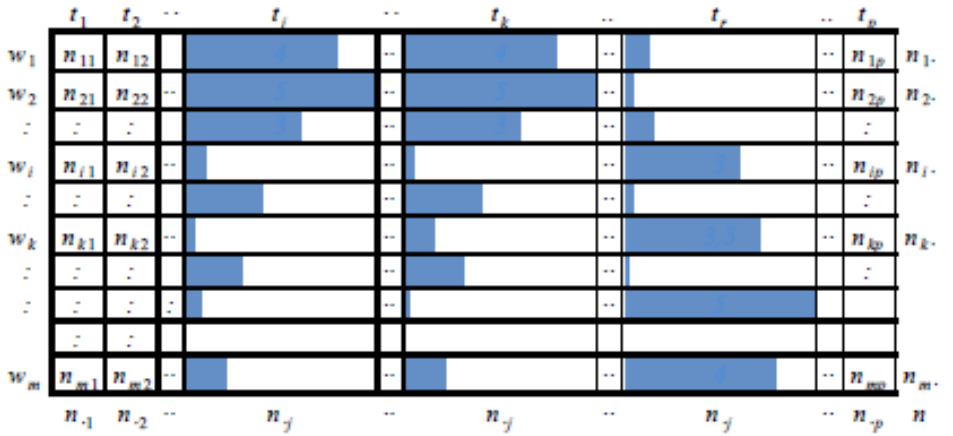
Distance between two texts t_j and t_l is given by:

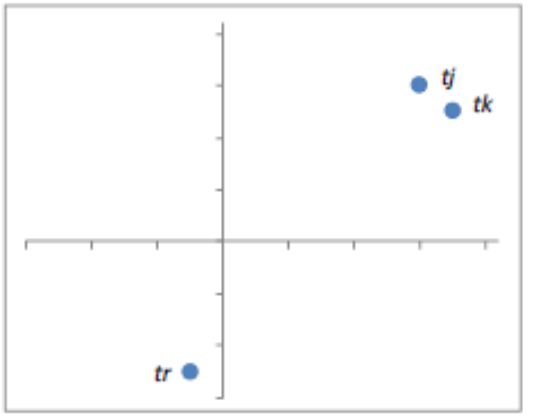
$$d^2_{jl(c)} = \sum_{i=1}^{I} \frac{1}{f_{i\bullet}} \left(\frac{n_{ij}}{n_{\bullet j}} - \frac{n_{il}}{n_{\bullet l}} \right)^2$$

where $f_{i\bullet} = \frac{n_{i\bullet}}{n}$ is the **mass** of row i (given by the marginal relative frequency of word i).

From an intuitive point of view it is easy to imagine that profiles in terms of relative frequencies can translate into near or far positions on a plane.

Example of profiles of three texts.





Inertia

The overall dispersion of each cloud of points around the respective centroid is called **total inertia**.

It is obtained as a weighted sum of the squared distances of the points from the respective centroid.

If we define the distance as the χ^2 metric, then we obtain that total inertia is:

$$\begin{aligned}\text{Inertia}_{\text{tot}} &= \sum_{i=1}^{m} r_i \sum_{j=1}^{p} \frac{1}{c_j}(x_{ij}^r - c_j)^2 \\ &= \frac{X^2}{n} = \phi\end{aligned}$$

The greater the total inertia is, i.e. the more the row or column profiles are distant from the respective centroids, the greater the association between the two categorical variables is.

On the contrary, the smaller it is, the greater the homogeneity between the profiles and their similarity to the average or marginal profiles is.

Interpretation of CA

- singular value decomposition (**svs**) of the matrix

$$E = R^{-\frac{1}{2}}(P - rc^{'})C^{-\frac{1}{2}} \quad \rightarrow \quad E = U\Gamma V$$

Interpretation of E: Matrix P (joint relative frequencies) is centered (i.e., is made independent from the centroid position) wrt both rows and columns as well as scaled (i.e., weight of categories is made independent from their mass).

Hence, a "**biplot**" of the matrix of row-point coordinates (G) and the matrix of column-point coordinates (H).

- metric multidimensional scaling (**MDS**) of the matrices of distance between row-points and column-points.

Hence, matrices of coordinates of row-points and column-points.

Two matrices of distances

After calculating the distances for all pairs of words and all pairs of texts, whence forming

- the square matrix `(m x m)` which contains the distances between the pairs of words
- the square matrix `(p x p)` with the distances in pairs between texts

the space generated by the original variables is transformed into an Euclidean space generated by new orthogonal variables (called components or axes).

The new variables are obtained by linear combination of the old variables and they are built in such a way as to be linearly independent from each other.

The multidimensional space generated by the original matrix is reduced to a space with orthogonal dimensions that can be represented with Cartesian axes.

The number of dimensions of this new space (the number of orthogonal axes) is equal to the number of linearly independent variables present in the data matrix (rank of the matrix).

In the case of a word matrix for texts with $m > p$ the number of axes is $p - 1$ (in general it is $\min(m, p) - 1$).

SVD

The calculation of the coordinates on each axis is based on the decomposition of matrix E into eigenvalues and eigenvectors called singular value decomposition (SVD).

The orthogonal factorial axes are ordered with respect to the amount of explained inertia (according to a logic of association), that is, they are set in order of importance: the first and most important axis is the one that gathers the highest portion of information contained in the contingency table, the second axis is the one that collects the highest portion of information not explained by the first axis and so on.

The Cartesian plane built with the first two factorial axes is the two-dimensional space which represents at best the association structure expressed by the contingency table.

Unlike other analyzes that move from a case $\times$ variable matrix, in CA the contingency table can be read in two ways:

- as m row-points (vectors) in the space generated by the columns
- as p column-points (vectors) in the space generated by the rows.

That premised, it is feasible to obtain two separate graphics: one with words and one with texts.

Duality

For the geometric properties of the two spaces (duality), the eigenvalues and eigenvectors are the same and the two graphs are superimposable.

In this way it is possible to observe the system of relations between all the categories involved; even if care must be taken in interpreting the joint graphic solution of the two variables.

Reading the graph

We briefly summarize some elements for reading the graphs obtained from CA.

1. the position of a word or text has a role only within the global context of the graph (that is, it makes no sense in itself, but makes sense in comparison with the positions taken by all the other points in the solution wrt the centroid located at the origin of the axes);

2. the words or texts that have contributed most to the solution and that, therefore, can be considered more important in the context reconstructed by the graph, are those far from the origin of the axes;

3. if two words are close they have similar (lexical) profiles;

4. if two texts are close they have similar (lexical) profiles;

5. the reciprocal position taken by a word and a text cannot be evaluated directly, because the distance between the two points on the graph does not translate automatically in terms of similarity: it must be evaluated with reference to the positions taken by all the other elements.

Note

With regard to point 5 it is useful to use the quadrants of the Cartesian plane because if a word and a text are placed on the same side with respect to the origin they can be considered positively associated (and negatively associated if they are in opposite positions).

Moreover, thanks to the axes, the proximity can be assessed taking into account the angle formed with the axes: the more the angle formed with the axes is similar, the more they can be considered associated.

Cluster search

Thickening of categories (words and texts) in an area of the graph that stands out from the remainder as a group (cluster) can be interpreted as a semantic area. This is why we often aim at obtaining a partition in clusters.

Word or text clusters must be as homogeneous as possible within group and as heterogeneous as possible between groups.

Building clusters through CA is just one of the many classification methods available.

CA in R

- ca (Nenadic and Greenacre)
- FactoMineR (Husson et al.)
- vegan (Dixon): for ecological data
- ade4 (Chessel): for ecological data
- anacor (de Leeuw and mair): simple and canonical CA
- homals (de Leeuw): multiple CA

Example

Suppose we have a small "toy" corpus consisting of 11 texts containing the keywords included in articles of a statistical journal.

text01	regression analysis; linear regression
text02	regression model; linear and non linear model
text03	generalized linear model; parameter estimation
text04	sampling methods; random sampling; survey design and sampling methods
text05	survey design; finite populations
text06	methods for sampling elusive populations
text07	Normal distribution
text08	z-scores and Normal distribution
text09	Gamma distribution
text10	p-value: Normal distribution and Gamma family
text11	regression analysis; Normal distribution

$N = 52$ word tokens

$V = 24$ word types

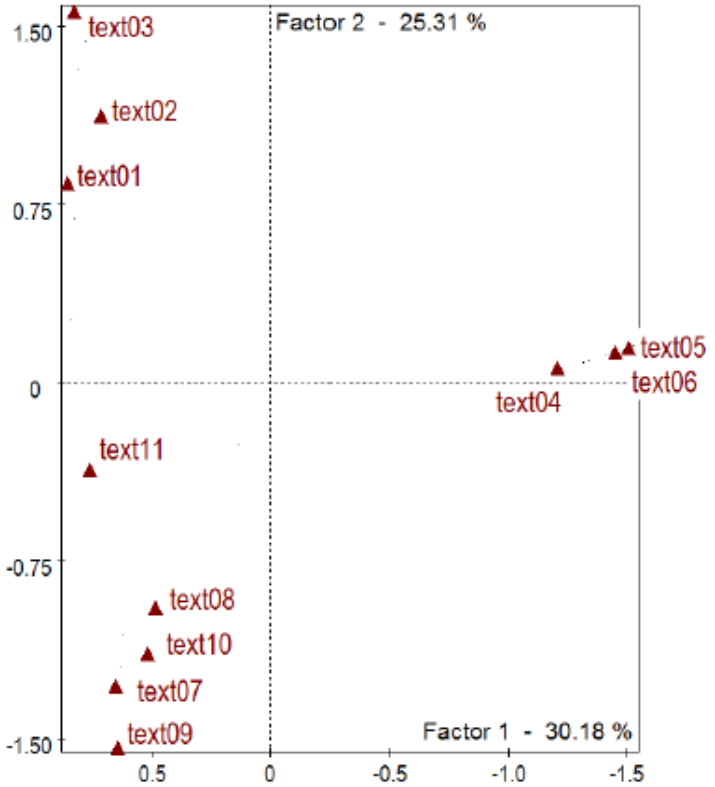
Example (continued)

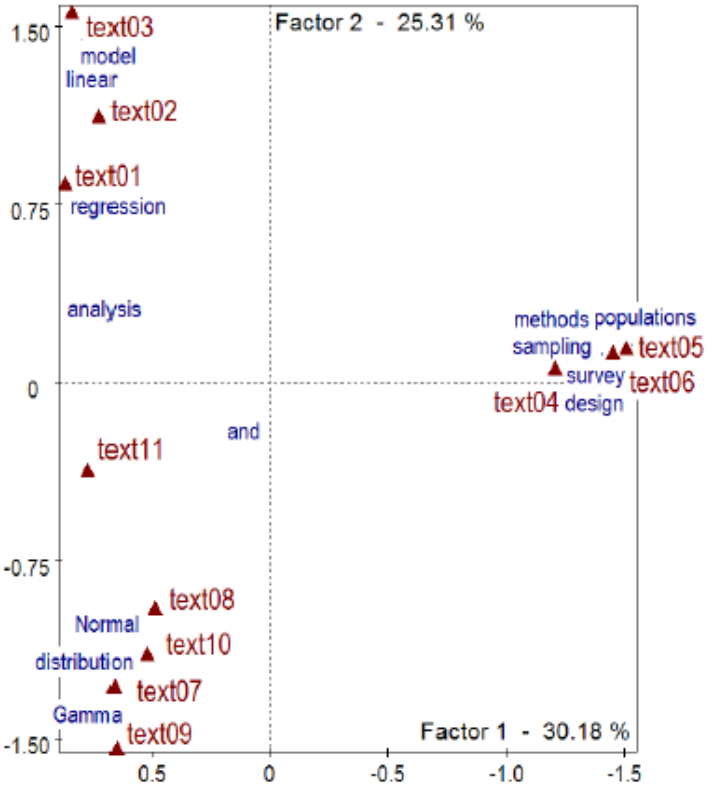
Taking into account only the words that repeat at least twice:

- *distribution*: 5 occurrences
- *and*, *linear*, *normal*, *regression*, *sampling*: 4 occurrences
- *metods*, *model*: 3 occurrences
- *analysis*, *design*, *Gamma*, *populations*, *survey*: 2 occurrences

words	text01	text02	text03	text04	text05	text06	text07	text08	text09	text10	text11
distribution	0	0	0	0	0	0	1	1	1	1	1
and	0	1	0	1	0	0	0	1	0	1	0
linear	1	2	1	0	0	0	0	0	0	0	0
Normal	0	0	0	0	0	0	1	1	0	1	1
regression	2	1	0	0	0	0	0	0	0	0	1
sampling	0	0	0	3	0	1	0	0	0	0	0
methods	0	0	0	2	0	1	0	0	0	0	0
model	0	2	1	0	0	0	0	0	0	0	0
analysis	1	0	0	0	0	0	0	0	0	0	1
design	0	0	0	1	1	0	0	0	0	0	0
Gamma	0	0	0	0	0	0	0	0	1	1	0
populations	0	0	0	0	1	1	0	0	0	0	0
survey	0	0	0	1	1	0	0	0	0	0	0

On the first Cartesian plane, texts, solely, or words and texts, jointly, can be displayed.





The lexical contingency table peoduces 10 orthogonal axes.

The first two axes collect 55% of information (explained **inertia**).

The figure clearly highlights three basic subjects/themes (latent patterns) in texts, which refer to

- linear models (linear, model, regression, analysis)
- sampling (sampling, methods, survey, design, populations)
- distributions (Normal, Gamma, distribution).

As regards texts

- texts 01, 02 and 03 are placed together in the area of linear models (second quadrant, top left)
- texts 07, 08, 09 and 10 in the distribution area (third quadrant, bottom left).
- text 11 is located between the area of linear models and that of distributions because it contains elements of both topics.
- texts 04, 05 and 06 are in the sampling area (first quadrant, on the right)

Finally, it is interesting to observe the position of the conjunction *and* which is close to the origin of the axes because it is used in different contexts (even if moved to the bottom left due to an extra occurrence in the group of texts referring to distributions).

Elena Ferrante

We have collected a corpus of contemporary Italian literature consisting of 150 novels (about 10 million occurrences) written by 40 different authors:

Affinati, Ammaniti, Bajani, Balzano, Baricco, Benni, Brizzi, Carofiglio, Covacich, De Luca, De Silva, Faletti, Ferrante, Fois, Giordano, Lagioia, Maraini, Mazzantini, Mazzucco, Milone, Montesano, Morazzoni, Murgia, Nesi, Nori, Parrella, Piccolo, Pincio, Prisco, Raimo, Ramondino, Rea, Scarpa, Sereni, Starnone, Tamaro, Valerio, Vasta, Veronesi, Vinci.

Novels and authors have been selected by several criteria:

1. novels by Elena Ferrante (published before 2018)

- L'amore molesto, Roma, E/O, 1992.
- I giorni dell'abbandono, Roma, E/O, 2002.
- La figlia oscura, Roma, E/O, 2006.
- L'amica geniale. Infanzia, adolescenza, Roma, E/O, 2011.
- Storia del nuovo cognome. L'amica geniale volume secondo, Roma, E/O, 2012.
- Storia di chi fugge e di chi resta. L'amica geniale volume terzo, Roma, E/O, 2013.
- Storia della bambina perduta. L'amica geniale volume quarto, Roma, E/O, 2014.

2. novels written by authors from the same geographical area (Naples / Campania);
3. novels written by authors suspected of being Elena Ferrante;
4. best sellers, award-winning novels (eg Strega);
5. novels written by authors who have met with critical acclaim

Corpus of 150 novels

- language: all novels are in Italian (original, untranslated)
- period: all novels were published in the period [1987-2016]
  - exceptions: Prisco 1966, Una spirale di nebbia; Prisco 1969, La provincia addormentata; Maraini 1972, Memorie di una ladra; Morazzoni 1986, La ragazza col turbante;
- Target: novels written for adult readers
- 13 women:
  - Ferrante (?), Maraini, Mazzantini, Mazzucco, Milone, Morazzoni, Murgia, Parrella, Ramondino, Sereni, Tamaro, Valerio, Vinci (50 novels)
- 11 authors from Campania:
  - Ferrante (?), De Luca, De Silva, Milone, Montesano, Parrella, Piccolo, Prisco, Ramondino, Rea, Starnone (46 novels)

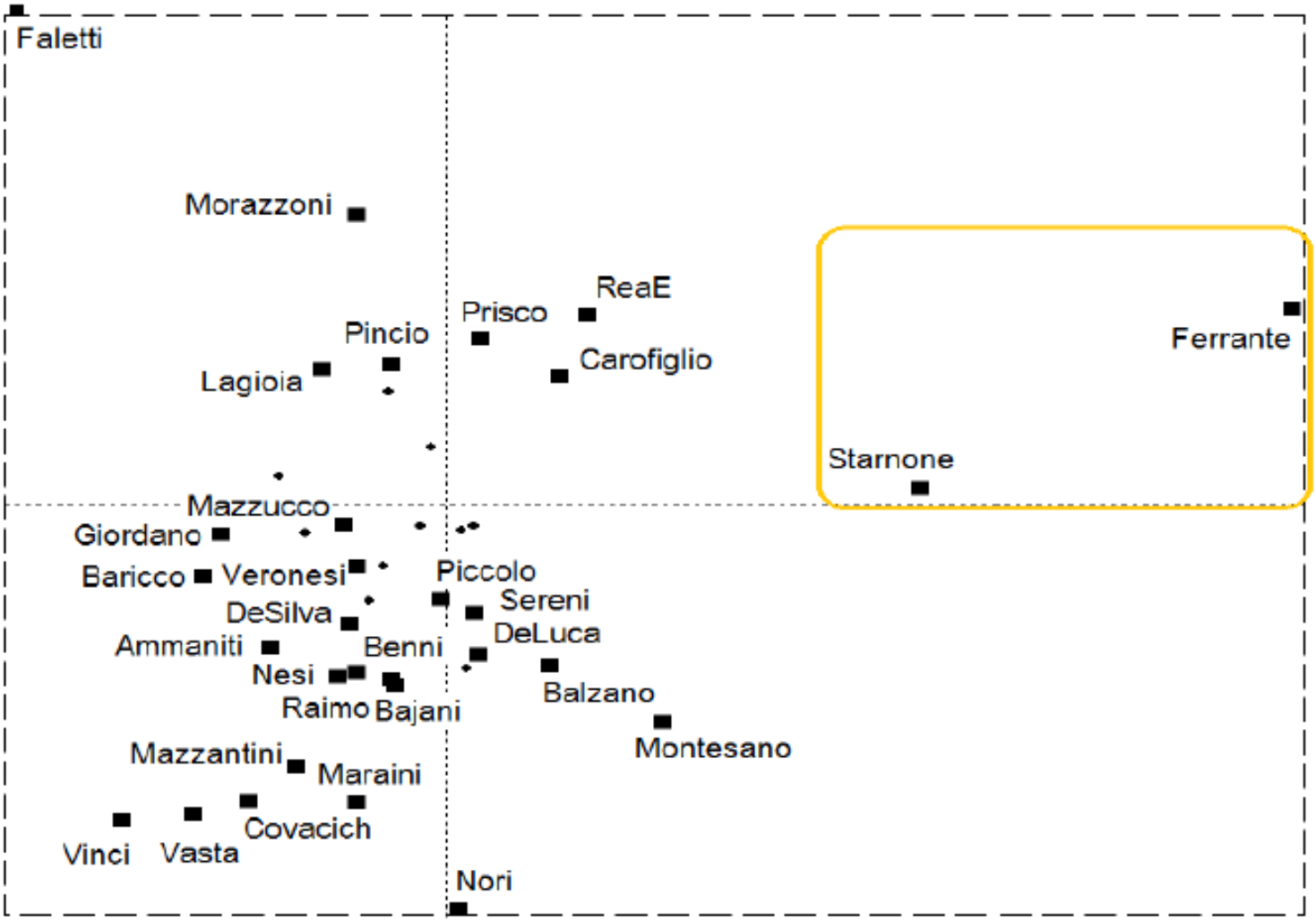
We consider the corpus organized by authors (subcorpora).

The corpus can be studied under different perspectives:

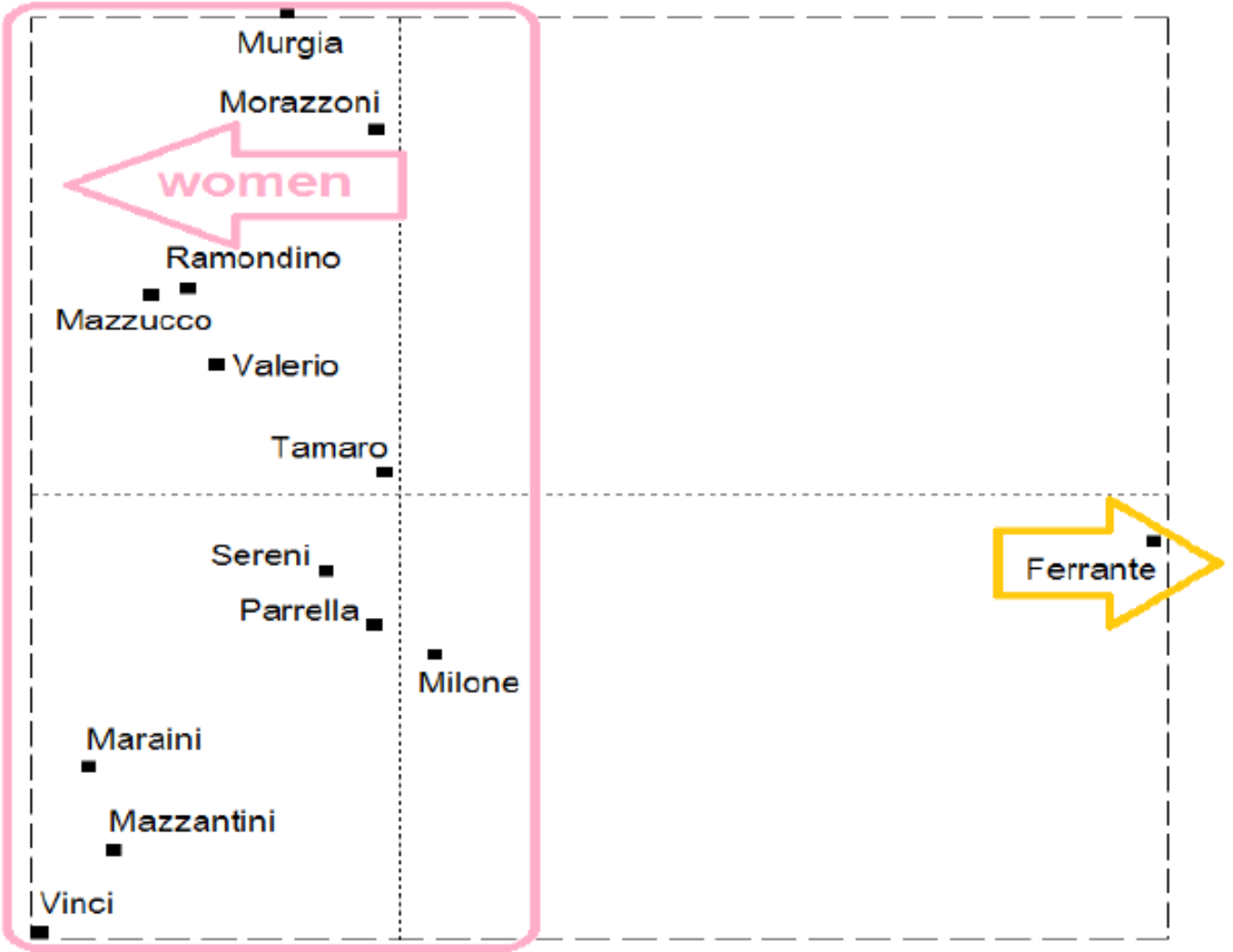
- fully (39 +1 authors);
- restricted to female authors (12 +1 authors)
- restricted to Neapolitan authors (11 +1 authors)

Where is Elena Ferrante? Who does it look like?

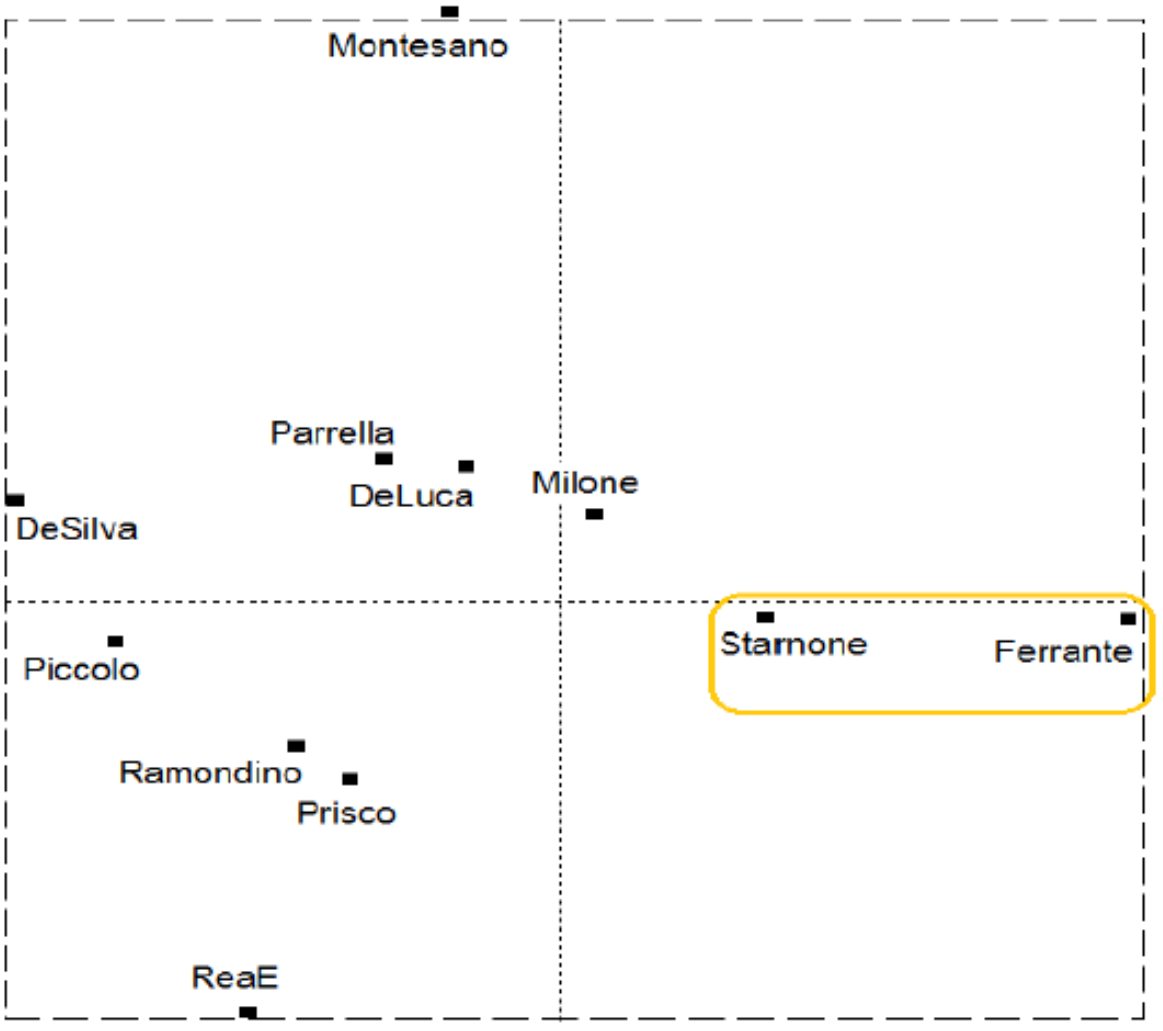
Elena Ferrante and 39 authors



Elena Ferrante and female authors



Elena Ferrante and authors from Napoli



Corpus of FC essays

In this case the corpus is small in size. It is:

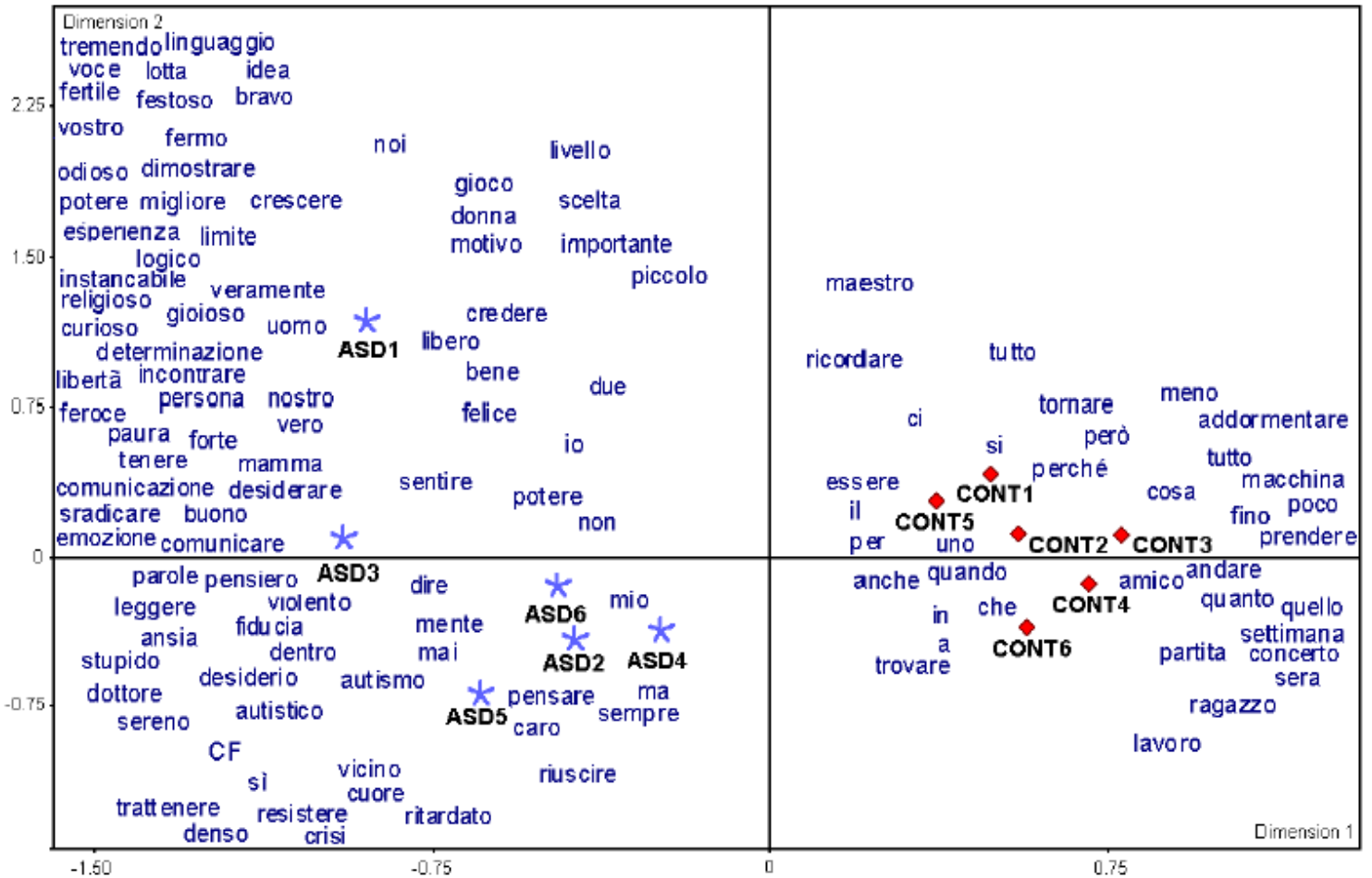
- a set of 12 essays
- produced within a case-control design
- from 6 couples (boys with autism / boys without disabilities of the same gender and age)
- during Facilitated Communication sessions

The title of the theme is

Describe a particularly significant day or moment for you

- **ASD**: case = boy / girl with autism (Autistic Spectrum Disorder)
- **CONT**: control case

Corpus of FC essays (continued)



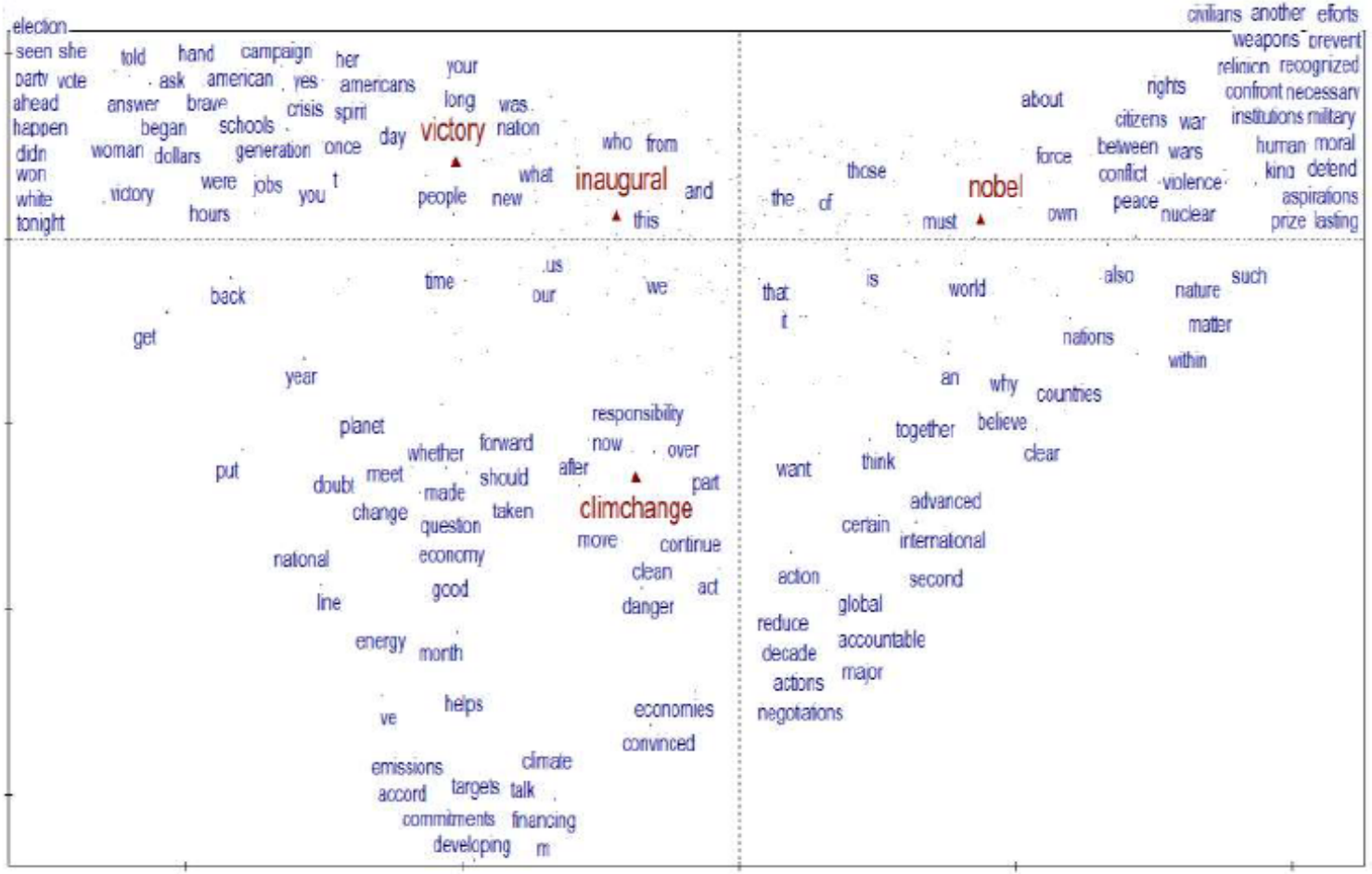
Obama

CA for 4 speeches from President Obama:

1. Presidential victory speech (Chicago, 4 novembre 2008)
2. Inaugural Address (Washington, 20 gennaio 2009)
3. Nobel Prize Lecture (Oslo, 10 dicembre 2009)
4. UN Climate Change Conference (Copenhagen, 18 dicembre 2009)

- 4 texts
- N = 9,783 word tokens
- V = 2,224 word types

Obama (continued)



78% explained inertia by first two dimensions