UNIVERSITÀ DEGLI STUDI DI PADOVA

Grammar and Lexicon in Texts Written by Individuals with Autism in FC Settings. Results from an Italian Interdisciplinary Research Program

Arjuna Tuzzi
arjuna.tuzzi@unipd.it

The EASIEST Project

[**E**spressione **A**utistica: **S**tudio **I**nterdisciplinare con **E**laborazione **S**tatistico-**T**estuale]

EASIEST is an Italian interdisciplinary research program funded by the University of Padua.

It involved scholars from different disciplines:

- linguistics and sociolinguistics
- sociology
- psychology
- psychiatry
- statistics

and FC-users and professionals of four Italian accredited FC-centers:

1. Padua
2. Zoagli (Genoa)
3. Rome
4. Andria (Bari)

Some methodological coordinates:

a) construction of a large database

b) three analytical approaches / three text corpora:

   b1) longitudinal on 13 FC-users with a long known history of FC training (400 pages, 130,000 words)

   b2) transversal on 37 FC-users who had reached a high level of independence with at least three different facilitators (900 pages, 290,000 words) → **Group2** corpus

   b3) experimental on 6 pairs "Cases vs Controls" (14 pages, 4,400 words) → **CaCo** corpus

Aims

1. identifying some lexical features of written communication of individuals with Autism
2. using these features for comparative purposes
   Group2: facilitators vs. users (*)
   Caco: cases vs. controls
3. identifying research issues that may be tackled integrating qualitative and quantitative methods
4. discussing what results may be obtained in an interdisciplinary environment

(*) important remark:
in our FC-sessions **facilitators type** questions and comments at the keyboard as well as their FC-users

Statistical Analysis of Textual Data

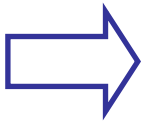
The field of Analysis of Textual Data (ATD) proves useful when studying any form of written (or transcribed) communication.

Which kind of texts?

In theory, any collection of texts. Examples in the educational field:

(a) transcriptions of oral conversations (interviews based on open-ended questions, case history and psychological interviews, educational talks)

(b) written production (diaries, essays, short stories, written open-ended answers)



(c) written conversations produced in FC settings

(d) utterances, documents, ethnographic notes, technical reports, press, public discourses, literary works, etc.

Background

A long tradition of studies…

- since its emergence as content analysis in the early 1900s, ATD has had a long tradition in the social sciences
- in linguistics, reference can be made to the great mathematicians of the past (Eustop, Mandelbrot, Marcov, Shannon, Zipf,…) and to the fathers of quantitative linguistics (Guiraud, Herdan, Muller, Reed, Yule,…)

Quantitative methods accounted for a niche within in the early 1970s.

Today the evolution of ICTs has led to many interrelated sectors: computational linguistics, quantitative linguistics, ATD, information retrieval, knowledge discovery, natural language processing, pattern recognition, machine learning, text mining…

It is a growing field!

Textual data

In terms of the textual data that is being analyzed, a traditional distinction may be made among:

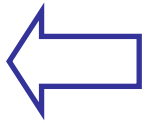
(a) phonetics

(b) grammar (morphology and syntax)

(c) lexis

Today we focus mainly on:

- lexis (**lexical analysis**)
- bag-of-words approaches



We deal with different textual units:

- form-types
- lemma-types
- multi-words

Rationale

- a **corpus** is a collection of texts
- a **text** is composed of blanks, letters, and other symbols (e.g., mathematical symbols, punctuation marks, numbers)
- a **word** is a sequence of letters isolated by means of separators (blanks and punctuation marks)
- in order to enumerate "the words" contained in a corpus we need to distinguish two concepts:

  word-tokens

  word-types

Types and tokens

A word-token is a particular occurrence of a word-type in a text

Example:
"the" is a word-type and has many tokens in any English text

(a) the number (N) of word-tokens is the **size of the corpus** in terms of occurrences

(b) the number (V) of word-types is the **size of the vocabulary** in terms of different words

(c) the **frequency** of a word-type is the number of corresponding word-tokens in the corpus

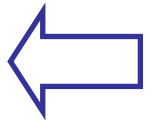
(d) the list of word-types and their frequencies is the **vocabulary** of the corpus

Lemmas

Lemmatization allows for analysis of the distribution of word-tokens among grammatical categories (parts-of-speech distribution).

Lemmatization is a **language-dependent concept** and plays different roles in different languages.

Texts written by individuals with Autism are very hard to lemmatize!



To analyze the word-types we can distinguish two concepts:

form-types

lemma-types

(a) form-types are the word-types "as they appear in the corpus"

(b) the lemmatization process associates each form-type with a pair that includes a lemma-type and a grammatical category

Examples:

went → go_V

students → student_N

Multi-words
(Segments, Compounds, n-grams…)

Words have different meanings if they are considered in their context of use and alongside the adjacent words.

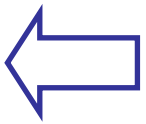
A **multi-word** is a sequence of words that appear many times in the same order (repeated segment).

Multi-words contribute to increase the amount of information conveyed by each word and prove useful for extracting high-quality information.

Examples:

mia cara Lisa [my dear Lisa]
furba Lisa [cunning Lisa]
molto bravo sono [clever very I am]
molto sono emozionato [moved deeply I am]
preside dell'istituto [dean of the institute-school]
mi stai prendendo in giro [you are kidding me]

With FC corpora a multi-words analysis often produced no significant results because the texts are seldom repetitive!



Two corpora

1. A **large corpus** including written conversations: **Group2**

- 38 individuals with Autism (FC-users)
- 92 facilitators
- 900 pages
- over 1,000 FC sessions
- 4 accredited Italian FC centers

The texts were produced during FC session and concern:

a) conversations in daily life

b) questioning about school-related experiences and topics

c) training interviews

d) text composition (essays, prose, etc.)

Two corpora

2. A small corpus including **12 short essays: CaCo**

- produced by six individuals with Autism (FC-users)
- and six participants in the control group (without disabilities)
- during sessions of FC
- using a case-control design (six pairs)
- using the same computer
- with the same facilitator
- with the same contact on his/her arms/shoulders
- at the same FC center (Padua, Italy)

Task:

writing a short essay of approximately 1.5 pages in length

Title:

Write about a moment/fact that was important to you

Size and lexical richness

Group2	N word-tokens	V form-types	TTR% (V/N)	H hapax	H% (H/V)
facilitators	159,243	12,359	7.8	6,301	51.0
IWA	131,253	14,875	11.3	8,373	56.3
corpus	290,496	20,166	6.9	10,055	49.9

CaCo	N word-tokens	V form-types	TTR% (V/N)	H hapax	H% (H/V)
Controls	2,145	793	37.0	524	66.0
IWA	2,215	981	44.3	688	70.1
corpus	4,360	1,550	35.6	1,039	67.0

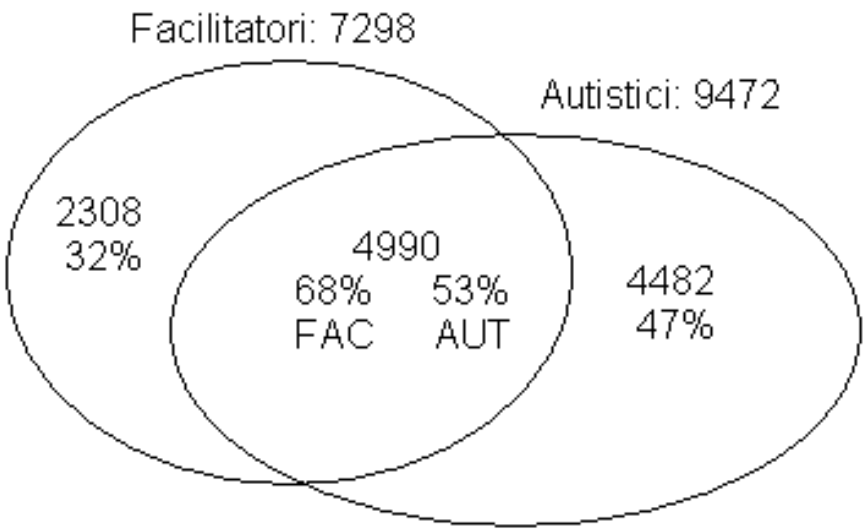
hapax legomena = word-type that occurs only once in the corpus/subcorpus

TTR = type-token ratio

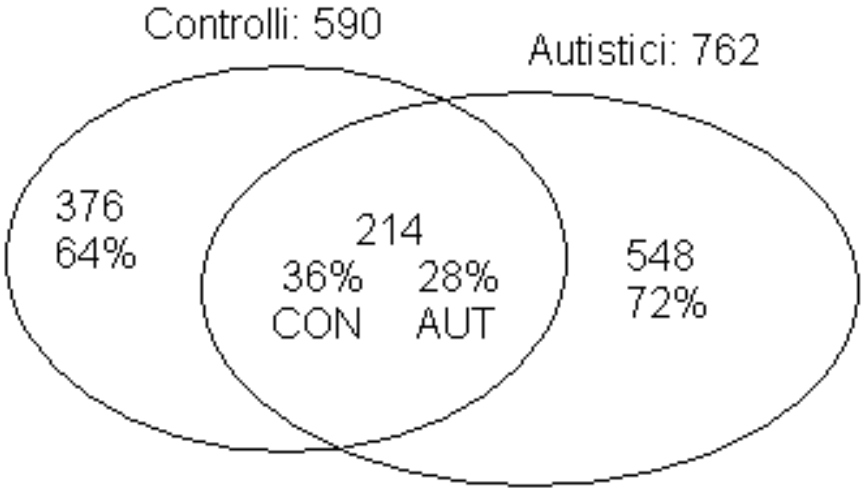
IWA = Individuals With Autism

Comparing and contrasting vocabularies

Group2

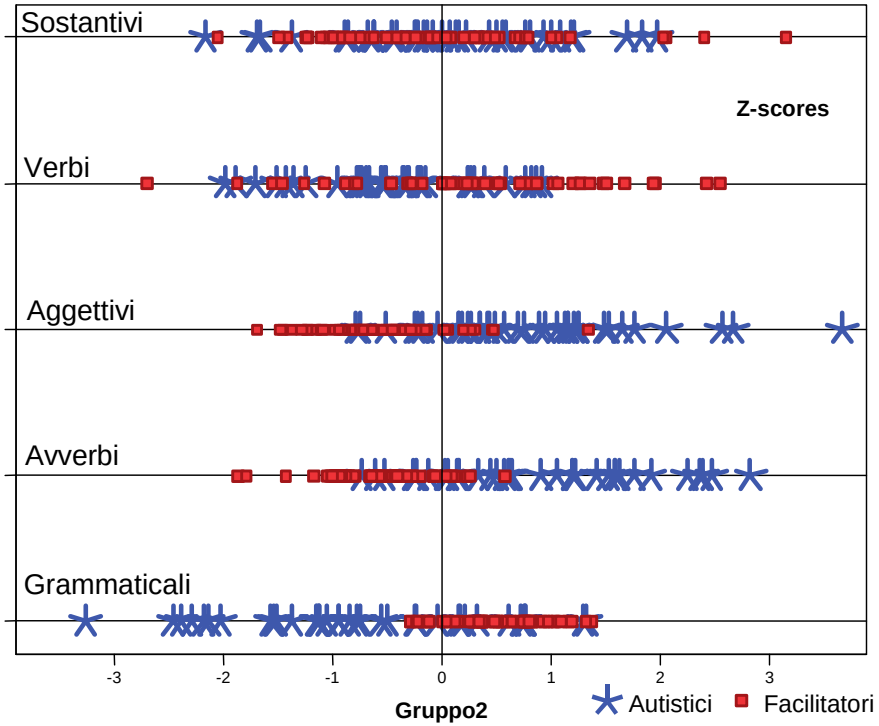


CaCo

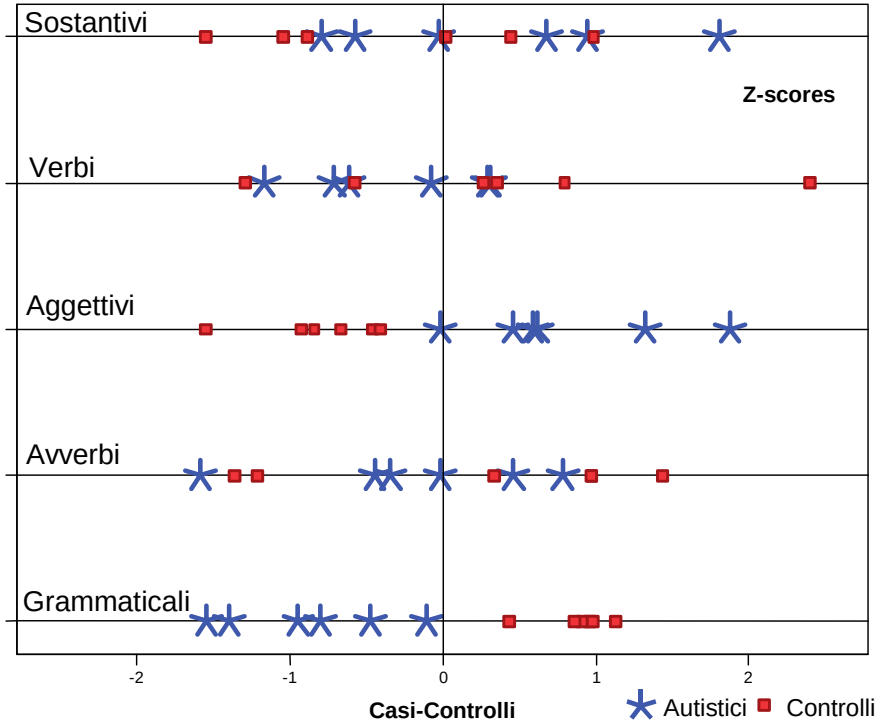


POS distribution

Group2



CaCo



Text clustering

Text clustering is a specific kind of unsupervised document classification concerning a corpus of texts in electronic format.

Text clustering is aimed at grouping similar texts to form consistent clusters and separating dissimilar texts into distinct clusters.

Text clustering involve several choices concerning:

1. which and how many words (→ **lemma-vocabulary**)
2. measures of (dis)similarity to be adopted (→ **intertextual distance**)
3. method to cluster texts (→agglomerative hierarchical cluster algorithm with **complete linkage**)
4. graphical representation (→ **dendrogram**)

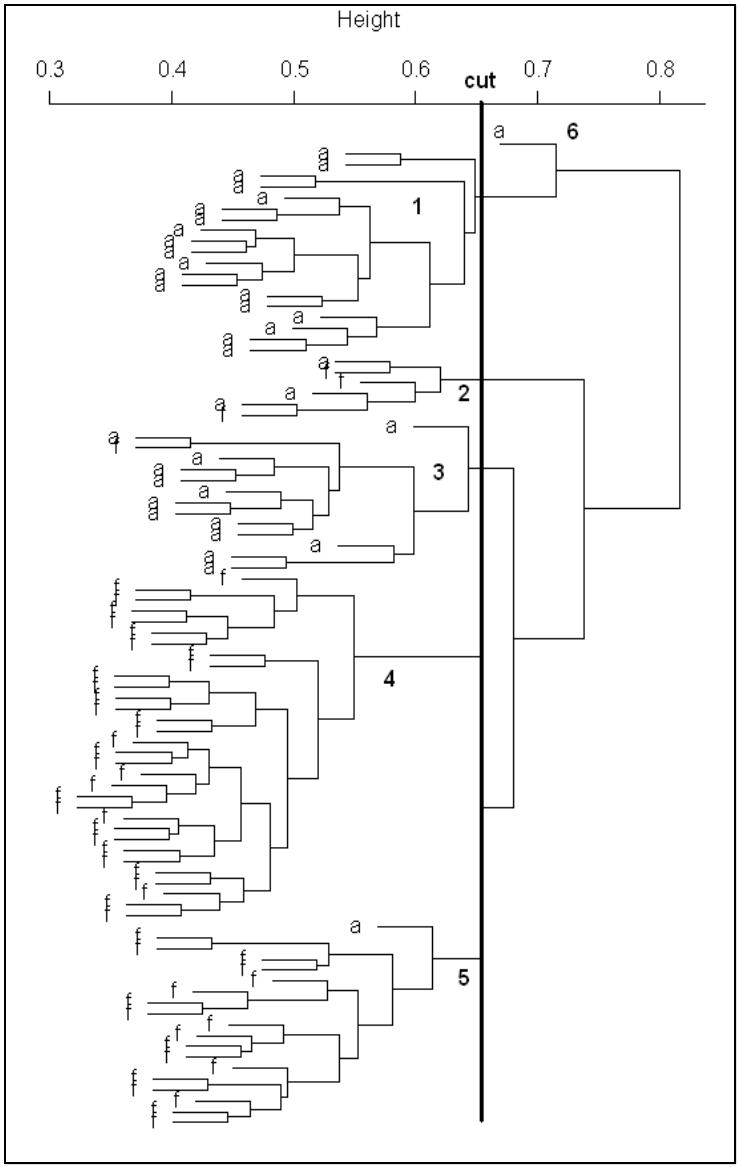
Some methodological notes:

- the list of lemma-types with the number of corresponding word-tokens mirrors the **lexical profile** of each text
- intertextual distance measures the pairwise **lexical distance** (sum of differences between the frequencies of words in texts A and B):

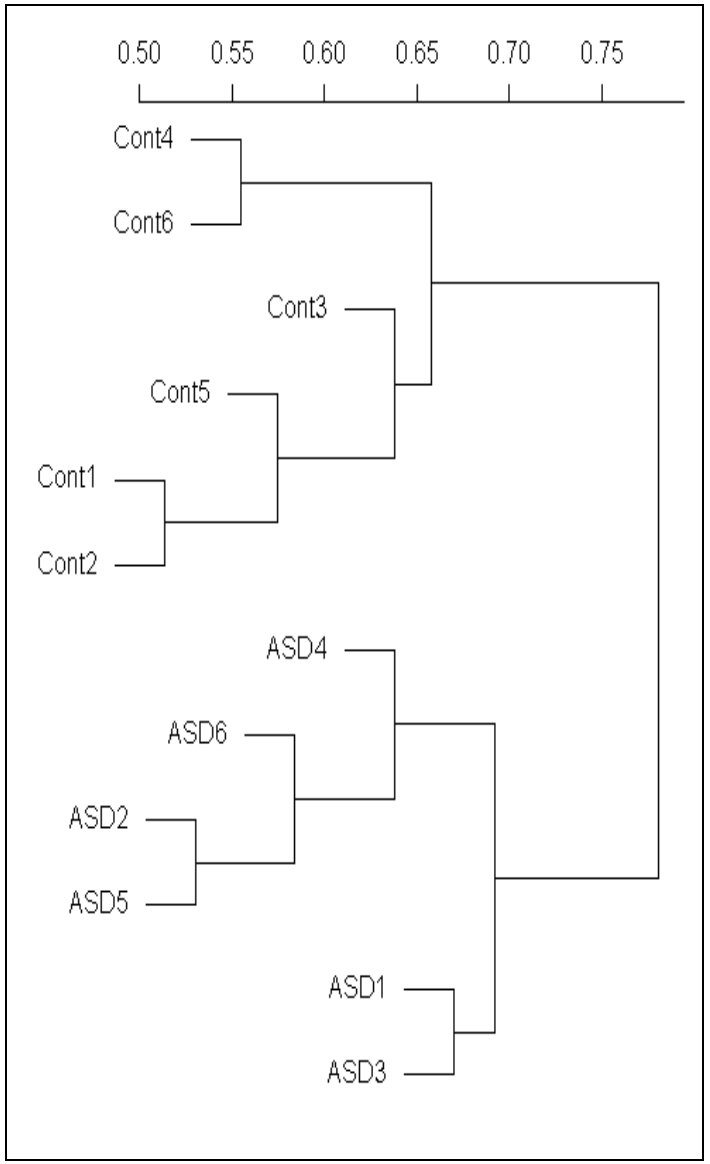
$$d(A,B) = \frac{\sum_{i \in V_{A \cup B}} \left| f_{i,A} - f^{*}_{i,B} \right|}{2N_A}$$

- the pairwise intertextual distances were evaluated through an agglomerative hierarchical cluster algorithm with **complete linkage** (the distance between pairs of clusters was obtained as the maximum distance among all pairs of elements of the two clusters and pairs of clusters with minimum distance were aggregated)
- **the dendrogram represents the distances by means of a tree graph in which the leaves (texts) hanging from the same branch form clusters of the closest texts and branches originating from the same forks represent groups of similar texts**

Group2


Height
0.3
0.4
0.5
0.6
cut
0.7
0.8
1
2
3
4
5
6

CaCo


0.50
0.55
0.60
0.65
0.70
0.75
Cont4
Cont6
Cont3
Cont5
Cont1
Cont2
ASD4
ASD6
ASD2
ASD5
ASD1
ASD3

Correspondence analysis

Correspondence Analysis is a multivariate explorative content analysis.

Correspondence Analysis is useful for highlighting the relations among authors, lemmas, and authors and lemmas.

Correspondence Analysis is aimed at transforming the lexical profiles of the authors in coordinates on a Cartesian axes system:

- the authors' positions on the plane are determined by the degree of similarity in terms of lexical profile (those who tend to use the same lemmas with the same frequency are close to each other and those who use different lemmas are positioned far from each other)
- **the authors who are plotted in the same quadrant of the plane deal with the same topics that are described by the lemmas of the same area of the plot**

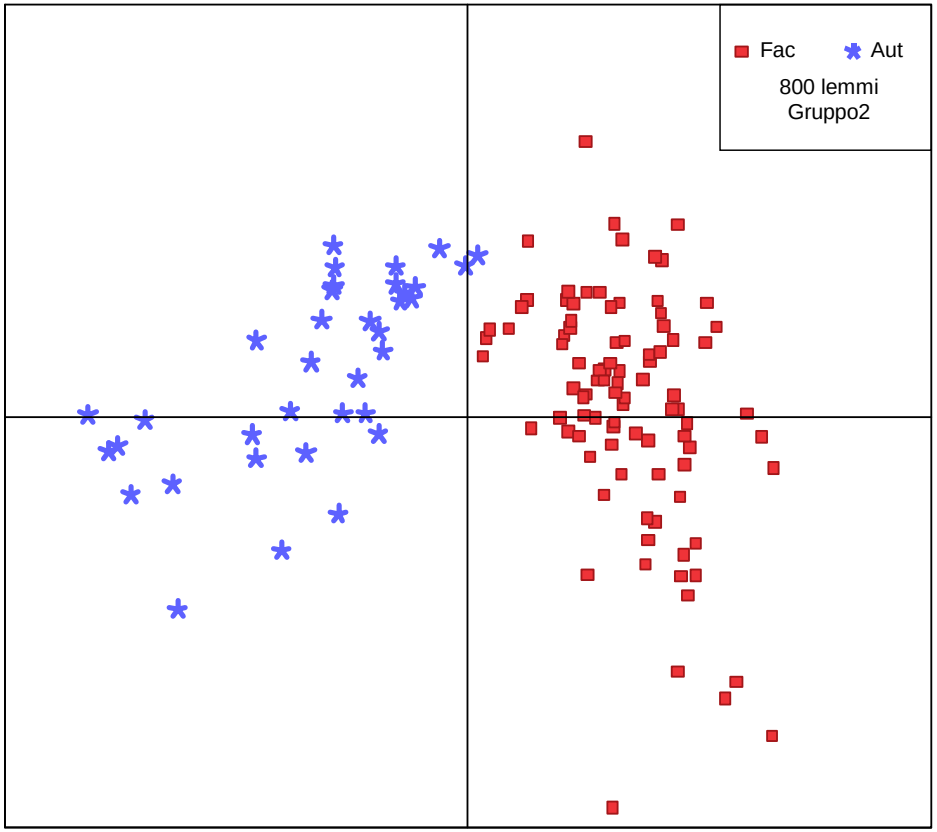
Some methodological notes:

- the list of lemma-types with the number of corresponding word-tokens mirrors the **lexical profile** of each text/author
- Correspondence Analysis is a special case of the Principal Component Analysis of the rows and columns of a table and the procedure is based on singular value decomposition (eigenvalues/eigenvectors)
- Correspondence Analysis displays the texts and the words in a low-dimensional space by mapping an appropriate distance (the chi-square distance) into a specific Euclidean distance and, then, into Cartesian planes
- finding theoretical interpretations for the extracted dimensions and the meaning of the axes is not the main goal
- **the following figures accounted for the lexical profiles of:**

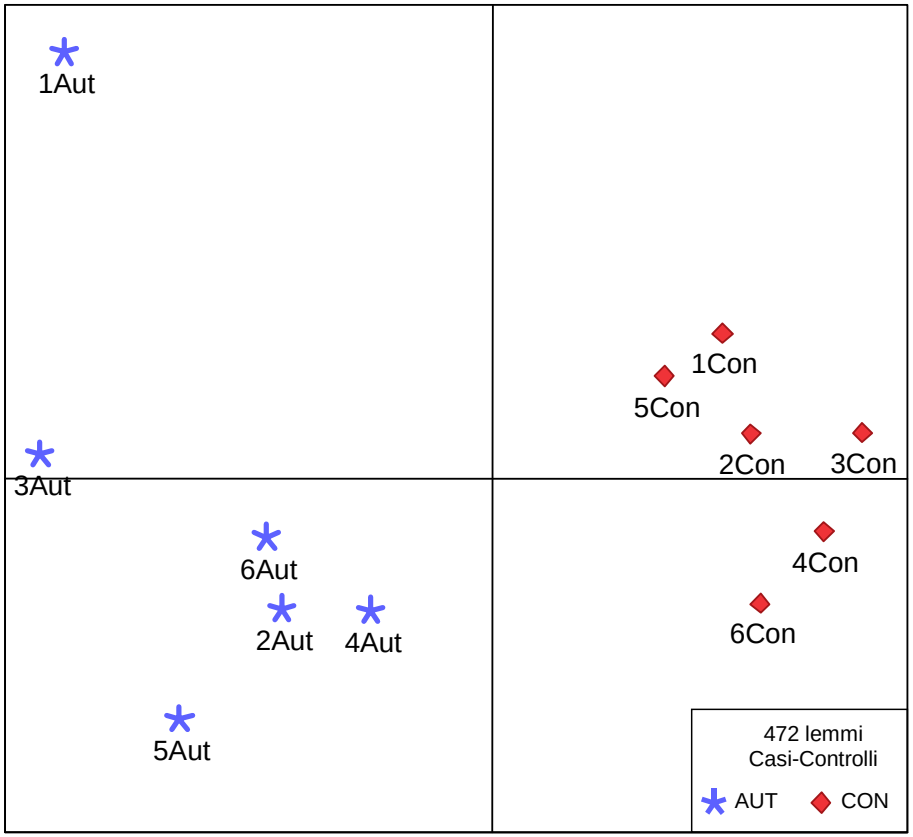
Group2 **37 + 92 FC-writers and 800 most frequent lemmas**

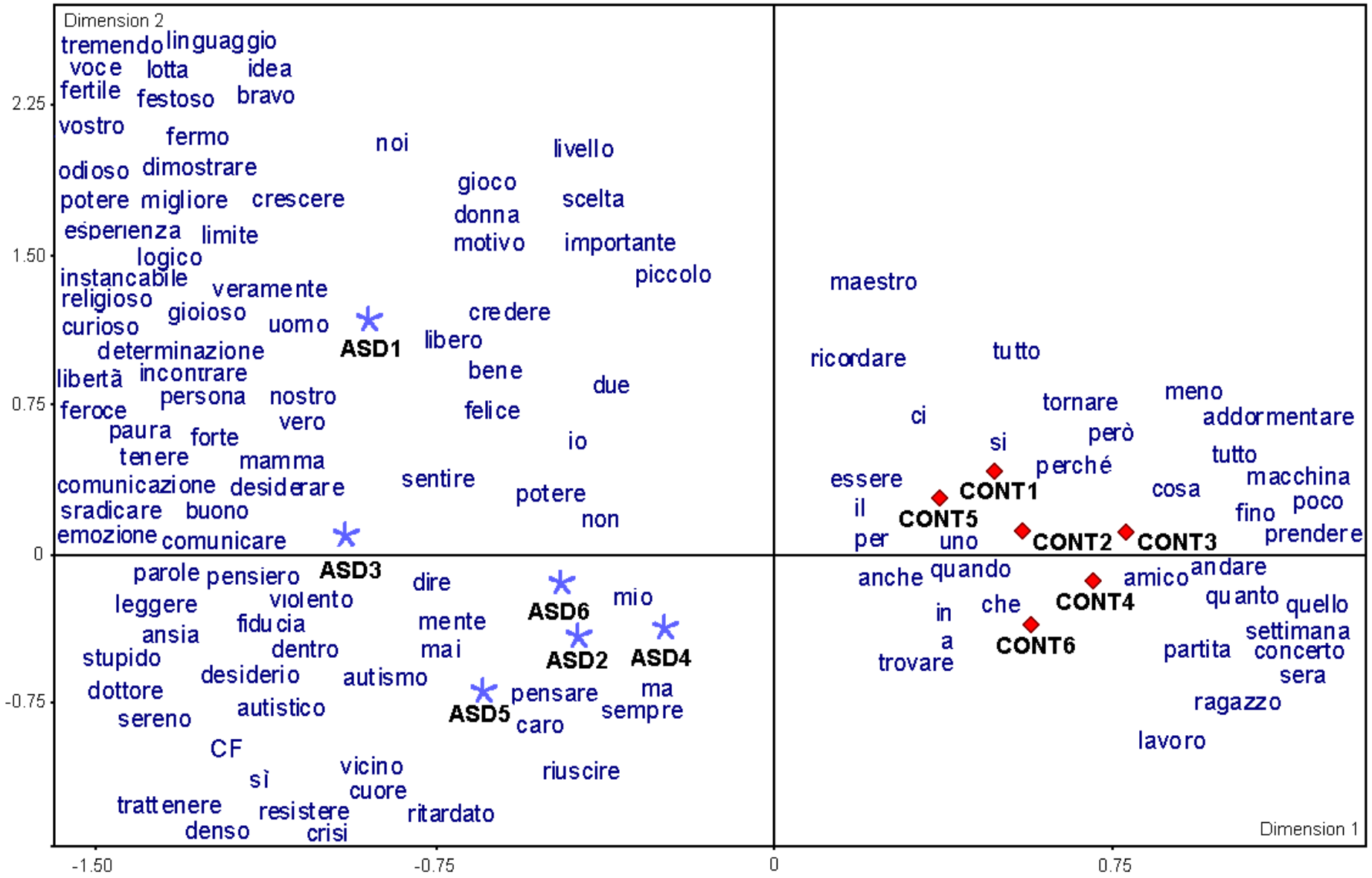
Caco **6 + 6 essays-authors and all 472 non-hapax lemmas**

Group2


Fac
Aut
800 lemmi
Gruppo2

CaCo


1Aut
3Aut
6Aut
2Aut
4Aut
5Aut
1Con
5Con
2Con
3Con
4Con
6Con
472 lemmi
Casi-Controlli
AUT
CON


Dimension 2
tremendo
linguaggio
voce
lotta
idea
fertile
festoso
bravo
vostro
fermo
noi
livello
odioso
dimostrare
gioco
potere
migliore
crescere
donna
scelta
esperienza
limite
motivo
importante
logico
piccolo
instancabile
veramente
maestro
religioso
gioioso
credere
curioso
uomo
ASD1
libero
determinazione
ricordare
tutto
bene
libertà
incontrare
due
persona
nostro
tornare
meno
ci
addormentare
feroce
felice
vero
però
si
paura
forte
io
tutto
tenere
mamma
perché
essere
macchina
sentire
CONT1
cosa
comunicazione
desiderare
potere
il
poco
CONT5
fino
sradicare
buono
non
per
uno
CONT2
CONT3
prendere
emozione
comunicare
parole
pensiero
ASD3
dire
anche
quando
amico
andare
ASD6
mio
leggere
violento
CONT4
quanto
quello
in
che
ansia
fiducia
mente
ASD2
ASD4
CONT6
settimana
a
stupido
dentro
mai
partita
concerto
trovare
desiderio
autismo
sera
dottore
pensare
ma
ASD5
ragazzo
sereno
autistico
sempre
caro
CF
lavoro
vicino
riuscire
sì
cuore
trattenere
resistere
ritardato
denso
crisi
Dimension 1
2.25
1.50
0.75
0
-0.75
-1.50
-0.75
0
0.75

Results

1. Lexical richness and POS distribution →

Findings:

the texts written by individuals with Autism showed a higher level of lexical richness, a greater proportion of adjectives and a tendency to omit grammatical words.

Implications:

the results support the hypothesis of the existence of **distinctive** lexical features and grammatical patterns.

Results

2. Text clustering and authorship attribution →

Findings:

the texts written by individuals with Autism are similar to each other and different from texts produced by other groups (facilitators and partecipants in the control group)

Implications:

the results support the hypothesis of the existence of **distinctive and consistent** lexical features

Results

3. Correspondence and content analysis →

Findings:

the contents expressed in texts written by individuals with Autism are similar to each other and deal mainly with feelings and emotions

Implications:

the results support the hypothesis of the existence of **distinctive, consistent and topic-consistent** lexical features

Discussion

Are these findings difficult to explain?

Yes, they are.

… but these phenomena exist and we have to deal with them

… in real life and with real (textual) data

The method/model must fit data, not vice versa

(I am a statistician)

ATD is a promising approach:

It poses great challenges and provides great opportunities in terms of applied research.

Interdisciplinarity

ATD is an interesting line of research because it deals with complex objects (corpora) and calls for expertise deriving from different disciplines.

An interdisciplinary expertise is essential to exploit (and appreciate?) the ATD approach.

Interdisciplinary research approaches are not always accepted by the scientific community and tend to play a marginal role as compared to more traditional single-discipline approaches...

EASIEST involved scholars from many disciplines: **linguistics** (Prof. Michele Cortelazzo; Prof. Flavia Ursini, Dr. Chiara Di Benedetto, Dr. Ivana Fratter), **neuropsychiatry and psychology** (Prof. Beatrice Benelli, Dr. Marisa Cemin, Dr. Vittoria Cristoferi), **sociology** (Prof. Federico Neresini, Dr. Stefano Sbalchiero), **statistics** (Prof. Lorenzo Bernardi, Prof. Arjuna Tuzzi, Dr. Alice Benato) and **professionals** and "**authors**" from the FC centers.

Qualitative or Quantitative?

ATD offers the opportunity to extract information from large amounts of complex, non-structured text data (systematically and time sparing)

ATD-Software provides numerous tools for the automatic and semi-automatic processing of data that are becoming increasingly powerful and continue to open new possibilities.

ATD overcomes the obstacles posed by the sheer size of corpora, which is one of the main limitations to qualitative analysis.

ATD is not meant to replace traditional qualitative approaches:

ATD should be seen as a type of integration rather than an alternative.

Size

Results become more reliable as corpora grow in size.

In order to analyse small texts, a traditional qualitative approach could be more suitable and powerful…

In the FC context few studies have focused directly on texts written during FC sessions, and few considered large corpora and several individuals.

Future work

In order to contribute to the scientific debate on the typical linguistic features of this specific written communication in general and the issue of authorship attribution in particular, you need:

- large corpora of texts produced during FC sessions
- written by many users
- with different facilitators
- in different settings
- (different languages?)
- …

Research questions:

- on the problem of using written language
- on the learning problem (chronological corpora)
- on the problem of the statistical methods adopted
- on the debate concerning the authenticity of texts generated using FC
- …

References:

Bernardi, L. (Ed.) (2008). Il delta dei significati. Rome, Italy: Carocci.

Tuzzi A., Cemin M., Castagna M. (2004), "Moved deeply I am" Autistic language in texts produced with FC. In: Purnelle G., Fairon C. e Dister A. (Eds), JADT 2004, Le poids des mots – Actes des 7èmes Journées internationales d'Analyse statistique des Données Textuelles, Loivain: UCL Presses universitaries de Loivain, 2, 1097-1105. [available on-line]

Tuzzi A. (2009). Grammar and Lexicon in Individuals With Autism: A Quantitative Analysis of a Large Italian Corpus. Intellectual and Developmental Disabilities, 47(5), 373-385.

Bernardi L., Tuzzi A. (2011), Analyzing Written Communication in AAC Contexts: A Statistical Perspective. Augmentative and Alternative Communication, 27(3), 183-194.

Bernardi L., Tuzzi A. (2011), Statistical Analysis of Textual Data from Corpora of Written Communication – New Results from an Italian Interdisciplinary Research Program (EASIEST). In: Mohammad-Reza Mohammadi (Ed.), A Comprehensive Book on Autism Spectrum Disorders, Rijeka: InTech, 413-434. [available on-line]

Grammar and Lexicon in Texts Written

by Individuals with Autism in FC Settings.

Results from an Italian Interdisciplinary Research Program

Thank you!