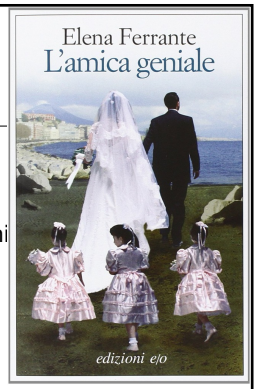
Who is the secret hand behind Elena Ferrante?

unine
UNIVERSITÉ DE NEUCHÂTEL

Jacques Savoy
University of Neuchâtel

1

In the pursuit of Elena…



Special thank to:

Prof. Michele Cortelazzo, Padova Uni
Prof. Arjuna Tuzzi, Padova Uni.

and other colleagues, and students

for generating the Ferrante corpus,
a corpus of high quality!

2

theguardian

Who is the real Italian novelist writing as Elena Ferrante?

As the fame of the Those Who Leave and Those Who Stay author grows, so does the guessing game about her identity

3

Elena F

End of author's anonymity

CULTURA | DOMENICA 2 OTTOBRE 201

Elena Ferrant

Toronto Star 1 Nov 2016 EMILY DONALDSON SPECIAL TO THE STAR

È una traduttrice di tedesco, e l'ipotesi circola da tempo: un'inchiesta di Claudio Gatti pubblicata dal Sole 24 Ore dice però di avere nuove prove

Ferrante fiever

Bestsellerautorin

Wer ist Elena F.?

Seit Jahren fragt die literarische Welt, wer sich hinter Italiens berühmtestem Pseudonym verbirgt: der Schriftstellerin Elena Ferrante. Hier kommt die Antwort

02.10.2016, von CLAUDIO GATTI

Photograph: Erich Lessing /

4

In the pursuit of Elena…

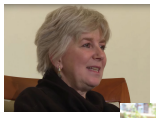
And in the Italian literature…

Who is “Elena Ferrante”?


Fiction

Elena Ferrante: the global literary sensation nobody knows

She shuns publicity and her identity is a mystery. Yet, as the last in her acclaimed series of novels about two friends in Naples is published, Elena Ferrante's reputation is soaring, with Zadie Smith, James Wood and Jhumpa Lahiri among her fans. Meghan O'Rourke on a literary mystery



Anita Raja?

Marcella Marmo?

Dominico Starnone?

And others.

5/22

5

Authorship Attribution

Two Noble Kinsmen

Shakespeare & Fletcher

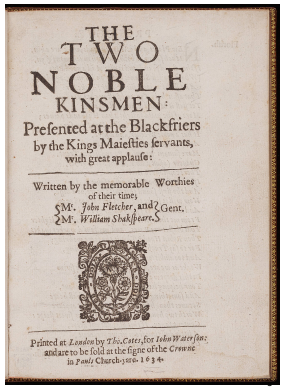
W. Shakespeare (1564-1616)

F. Bacon (1561-1626)?

C. Marlowe (1564-1593)?

E. De Vere (1550-1604)?
Lord Oxford

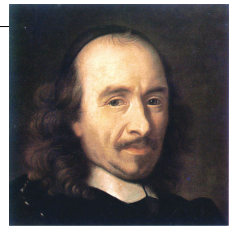
Giovanni Florio (1553-1625)?


THE TWO NOBLE KINSMEN:
Presented at the Blackfriers by the Kings Maiesties servants, with great applause:
Written by the memorable Worthies of their time; M. John Fletcher, and Gent. M. William Shakspeare.
Printed at London by Tho. Cotes, for John Waterson; and are to be sold at the signe of the Crowne in Paul's Church-yard. 1634.

6

Authorship Attribution

- The debate *Molière* vs. *Corneille*?
  Jean Baptiste Poquelin (1622-1673)
  Pierre Corneille (1606-1684)
- *Psyché* (1671), both are authors



Labbé, D. (2009). *Si deux et deux font quatre, Molière n´a pas écrit Dom Juan*. Paris, Max Milo.

7

Gary – Ajar in France

- Romain Gary (1918-1980)
  French novelist.
- Emile Ajar appears in 1973 with the novel *Gros-Câlin*
  “A real new style” said the press.
- 1974: starting to look for the real name behind “Ajar”.
- 1975: Gary’s cousin is the “Ajar”?
- 1980: R. Gary wrote that he is the real E. Ajar.

8

Galbraith–Rowling Case

- Since 2007, it seems that J.K. Rowling will write a crime novel.
- *The Cuckoo's Calling* appears in April 4th, 2013 under the penname Robert Galbraith.
- Who is this R. Galbraith?
- July 13th, 2013, *The Sunday Times* indicates that J.K. Rowling was the author of this novel.

9

Authorship Attribution

- Who wrote this text?
- Is this document a copy of another?
- Was this poem written by Shakespeare?
- Is the style stable over the author's life?
- How many writers can be found in this set of papers?
- Is this novel written by a woman?
- Identify in this collaborative text, passages written by each author

10

Variations in Style

What do you mean by *style*?

1. The village does not have a post office.
2. The village has no post office.
3. The village doesn't have a post office.
4. The village hasn't got a post office.
5. The village hasn't got no post office.
6. The village ain't got no post office.

Personal choice (aesthetic reasons)

Crystal, D. (2010). *A Little Book of Language*. Yale University Press

11

11

Stylistic Components

- Style is a function of
  - **Genre** (novel vs. poem, prose or verse)
  - **Author** (social, gender, age, education, native language, …)
  - **Period** (each period has its own stylistic preference)
  - **Topic**
  - Type (spoken vs. written, web-based)
  - Audience (official vs. informal)
  - Editors / publishers

12

Text is a Composite Signal

Text genre
Period
Topic
Type
Audience
Author
gender
age
psychological traits
social class
native language
education

13

13

Qualité des données…

Because data quality matters!

FC Bosna

fc bosna j aisspaire que momo revienne sest le seule qui peux nous faire monté comme entreneure

Toi t'est un qui n'a rien vue

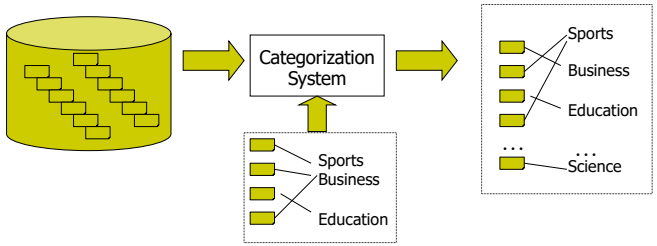
Je te rappel que contre le fcc sur 11 joueur sur le terrain il y avait que 4 joueurs licencié. Sinon c'était des essais pour voir le niveau des nouveaux arrivées.

14

14

Text Categorisation

- Predefined categories C (categories may form hierarchy)
- Set of labeled document examples D (to learn)
- A standard classification (supervised learning) problem



15

Example: Written by ?

A man or a woman?

and why?

```
Yesterday we had our second jazz
competition. Thank God we weren't
competing. We were sooo bad. Like, I
was so ashamed, I didn't even want to
talk to anyone after. I felt so
rotton, and I wanted to cry,
but...it's ok.
```

16

16

Example: Written by ?

A man or a woman?
and why?

```
My gracious boss had agreed to let me
have one week off of "work." He did
finally give me my report back after
eight freakin' days! Now I only have
the rest of this week and then one
full week after my vacation to finish
this damned thing.
```

17

17

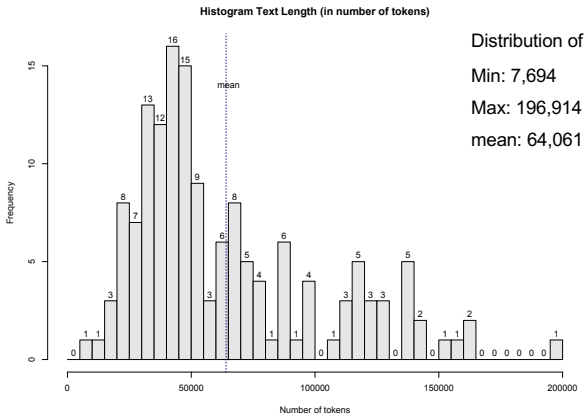
Ferrante corpus

- 150 novels written by 40 authors
- 10 from *Napoli*, 12 women

Affinati 2	*De Silva* 5	*Montesano* 2	*Ramondino* 2
Ammaniti 4	Faletti 5	Morazzoni 2	*Rea* 3
Bajani 3	Ferrante 7	Murgia 5	Scarpa 4
Balzano 2	Fois 3	Nesi 3	Sereni 6
Baricco 4	Giodano 3	Nori 3	*Starnone* 10
Benni 3	Lagiola 3	*Parrella* 2	Tamaro 5
Brizzi 3	Maraini 5	*Piccolo* 7	Valerio 3
Carofiglio 9	Mazzantini 4	Pincio 3	Vasta 2
Covacich 2	Mazzucco 5	*Prisco* 2	Veronesi 4
De Luca 4	*Milone* 2	Raimo 2	Vinci 2

18

In the pursuit of Elena…



Distribution of the text length

Min: 7,694

Max: 196,914

mean: 64,061

19

In the pursuit of Elena…

- What are the possible features?
  (Build a language model per author)
- How can we select the best subset?
- What are the good classifiers (distance-based methods)?
- And finally, can you explain why?

20

In the pursuit of Elena…

- Very frequent tokens (lemmas), functional terms
- A subset / entire vocabulary
  Feature selection function: odd ratio, info. gain, …
- Letter frequencies, unigram / bigrams / letter *n*-grams
- Topical terms (unigram, bigram, *n*-gram)
- Syntactical variables
  - POS distribution, sequence of POS tags
  - Mean sentence length (MSL)
- Metadata information (action in Napoli, in 50s)

- Approved attribution methods

21

Comparative Stylistic Study

	Washington		Lincoln		Obama	
Rank	**Lemma**	**Freq.**	**Lemma**	**Freq.**	**Lemma**	**Freq.**
1	the	1,545	the	2,426	the	2,084
2	of	1,098	of	1,485	*we*	2,036
3	to	669	be	1,104	be	1,633
4	be	665	and	952	and	1,632
5	and	517	to	856	to	1,604
6	an	370	in	535	of	1,164
7	in	292	an	436	an	1,077
8	have	276	have	419	that	1,042
9	which	222	it	391	in	757
10	they	203	that	314	have	635

8 lemmas in common!

22

Style and Function Words

Lemma	Carofiglio	De Luca	Ferrante	Starnone
il	4,18 %	**5,86 %**	**4,33 %**	4,55 %
di	2,82 %	2,80 %	2,55 %	2,56 %
e	2,43 %	2,31 %	2,13 %	2,14 %
essere	**2,65 %**	2,11 %	**2,07 %**	2,06 %
che	2,15 %	1,59 %	2,36 %	2,10 %
a	1,44 %	1,87 %	1,92 %	1,65 %
avere	1,73 %	1,11 %	1,54 %	1,70 %
un	1,52 %	1,60 %	1,12 %	1,25 %
del	1,25 %	1,31 %	1,10 %	1,21 %
non	1,52 %	1,47 %	1,42 %	1,23 %

23

23

Evidence #1: Labbé's Distance

Intertextual distance

Using the entire vocabulary
Compare the term frequency (*tf*) in Text A and Text B
(min value: 0, max value: 1)

$$Dist_{Labbe}(A,B) = \frac{\sum_{i=1}^{m} |tf_{iA} - \hat{tf}_{iB}|}{2 \cdot n_A}$$

$$with \quad \hat{tf}_{iB} = tf_{iB} \cdot \frac{n_A}{n_B}$$

24

Intertextual distance (Labbé, 2007)

Two texts with the different sizes ($n_A = 4$, $n_B = 8$)

Text A
Yes, we can scan.

Text B
Yes, we can, and we can do more.

tf^A	tf^B	tf^{B*}
yes: 1	yes: 1	yes: 0.5
we: 1	we: 2	we: 1
can: 1	can: 2	can: 1
scan: 1	and: 1	and: 0.5
	do: 1	do: 0.5
	more: 1	more: 0.5

$D_{rel}(A,B) = (0.5+0+0+1+0.5+0.5+0.5) / (2 \cdot 4) = 3 / 8 = 0.375$

25

Evidence #1

- Each novel is treated separately
- Represent each novel by its tokens / lemmas
- Selected all tokens if *tf* > 2 (reduce the voc. size of 50%)

ID	Year	Title
46	1992	L'amore molesto
47	2002	I giorni dell abbandono
48	2006	La figlia oscura
49	2011	*L'amica geniale*
50	2012	*Storia del nuovo cognome*
51	2013	*Storia di chi fugge e di chi resta*
52	2014	*Storia della bambina perduta*

26

Distance between Novels (lemma)

Rank	Distance	ID	Name	ID	Name	
1	**0.111**	51	Ferrante	52	Ferrante	
2	**0.121**	50	Ferrante	51	Ferrante	
3	**0.128**	49	Ferrante	50	Ferrante	
4	**0.134**	50	Ferrante	52	Ferrante	
5	**0.142**	145	Veronesi	147	Veronesi	
6	**0.146**	42	Faletti	44	Faletti	
7	**0.150**	43	Faletti	44	Faletti	
8	**0.154**	41	Faletti	42	Faletti	
9	**0.157**	42	Faletti	43	Faletti	
10	**0.161**	38	De Silva	39	De Silva	
11	**0.161**	49	Ferrante	51	Ferrante	
…						
33	**0.193**	52	Ferrante	132	Starnone	<-
38	**0.195**	51	Ferrante	131	Starnone	<-
41	**0.196**	51	Ferrante	132	Starnone	<-
…						
84	**0.216**	25	Carofiglio	147	Veronesi	Fails

27

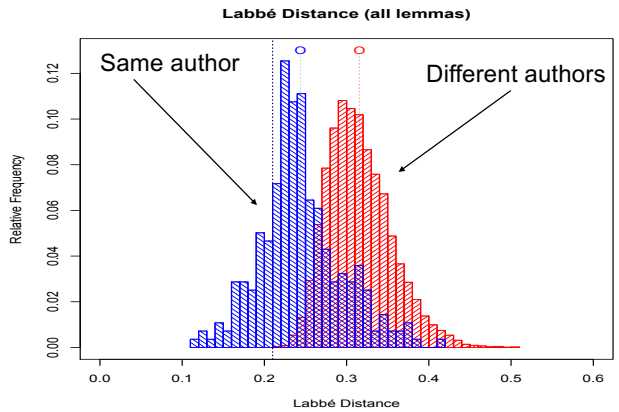
Too Small Distance

Year	Title	Author	Year	Title	Author
2014	Storia Della Bambina Perduta	Ferrante	2014	Lacci	Starnone
2002	I Giorni Dell abbandono	Ferrante	2007	Prima esecuzione	Starnone
1992	L'amore molesto	Ferrante	1993	Eccesso di zelo	Starnone
2013	Storia di chi fugge e di chi resta	Ferrante	2011	Autobiografia erotica …	Starnone
2011	L'amica geniale	Ferrante	2000	Via Gemito	Starnone

Words used only by Starnone & Ferrante:
(uccisore, studenti, argentiera, immaginai, confrontabile, calcolatamente, …)

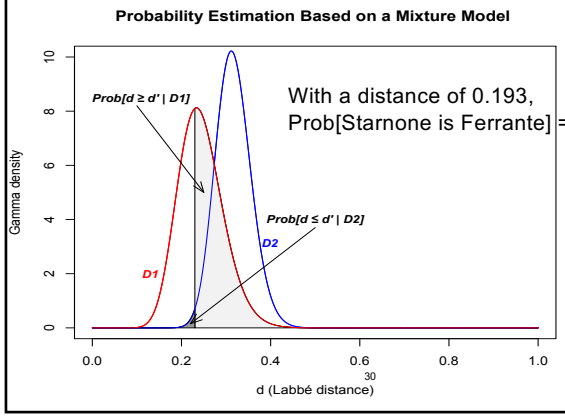
28

Mixture of two distributions



29

Mixture of two distributions



With a distance of 0.193,
Prob[Starnone is Ferrante] = 0.98

Savoy, J. (2016). Estimating the Probability of an Authorship Attribution. *JASIST*, 67(6), 1462-1472.

30

But models could be limited…

The FIFA World Ranking 2019

Rank	Country	Country
1	Brazil	Belgium
2	Germany	France
3	Argentina	Brazil
4	Switzerland	England
5	Poland	Portugal
6	Portugal	Uruguay
7	Chile	Spain
8	Columbia	Croatia
9	Belgium	Colombia
10	France	Argentina

31

Evidence #2

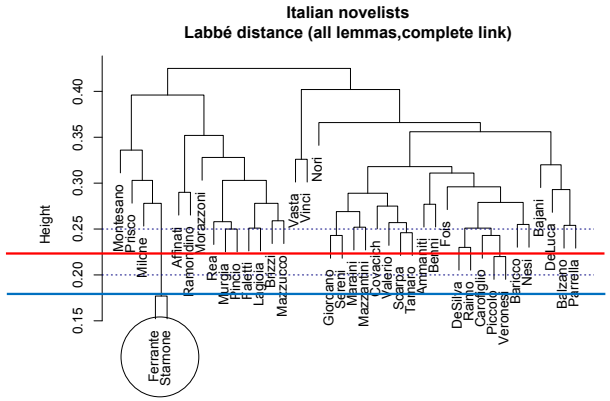
- Concatenated all novels written by the same author.
- Represent each author by the lemmas (*tf*>2).
- Apply the Labbé's distance (min: 0; max: 1).

Conclusion

- Min distance over all pairs of authors = 0.177
  This min value is between Ferrante & Starnone.
- Dist (Ferrante, Starnone) = 0.177
  Dist (Picollo, Veronesi) = 0.22
  Dist (Nesi, Veronesi) = 0.226
  Dist (De Silva, Veronesi) = 0.227

32

Distance (Labbé)



33

Evidence #3: Zeta Test

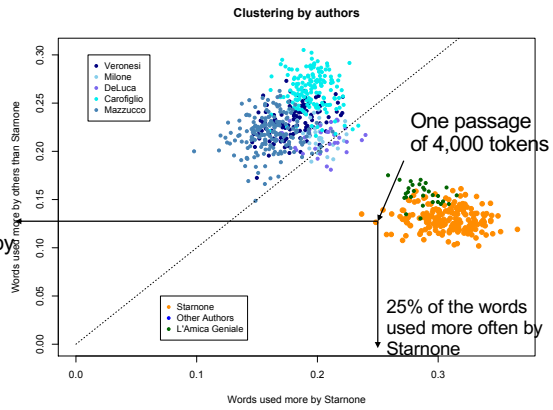
- Identify words used more frequently by Starnone
  (e.g., *volta, er, sua, solo, poi, senza, della, quando*, …)
  and words used more frequently by others (but not Ferrante)
  (e.g., *volto, muro, strana, nonstante, propria, ognuno*, …)
- The percentage of words present (used by Starnone or not) in a passage of 4,000 tokens corresponds to both axis values

Conclusion

- Example*: Amica Geniale* & Starnone
  cluster of points on the same region
  and distinct from the others

34

Zeta test / Evidence #3



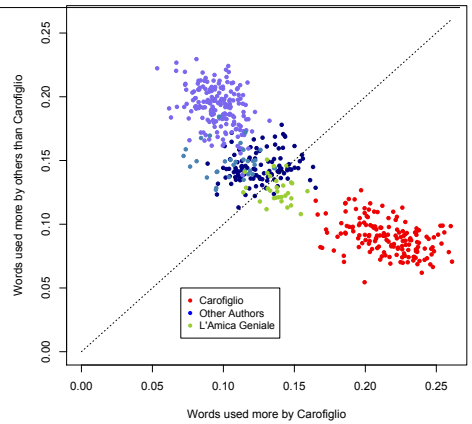
35

Zeta Test

Counter-example

G. Carofiglio

Different from E. Ferrante



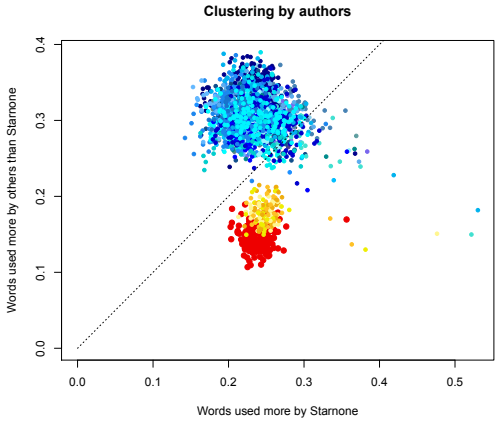
36

Zeta Test

With 39 authors and 144 novels

And the novel *L'amica geniale* In yellow…

Starnone in red The others in blue.



37

Evidence #4: Burrows' Delta

- Based on on the *m* most (*m* = 50 to 1,000) frequent words
- Compute a Z-score value for each word
  - for each word type w_{ij} , *i* = 1, …, *m* in a document D_j, compute the relative frequency rf_{ij} (in ‰)
  - μ_i mean of *i*th word-type in the reference corpus
  - σ_i standard deviation

$$Z\ score(w_{ij}) = \frac{rf_{ij} - \mu_i}{\sigma_i}$$

Burrows, J. F. (2002). Delta: A measure of stylistic difference and a guide to likely authorship. *Literary and Linguistic Computing*, 17(3), 267-287.

38

38

Burrows' Delta

Then the relative frequencies, and the mean and stdev.

	H59	H60	H61	H62	M37	M38	M47	M48	μ	σ
the	0.258	0.272	0.273	0.295	0.239	0.236	0.307	0.252	0.267	0.024
,	0.194	0.185	0.187	0.179	0.200	0.202	0.205	0.236	0.199	0.017
of	0.164	0.176	0.180	0.174	0.165	0.163	0.175	0.148	0.168	0.010
to	0.107	0.106	0.110	0.112	0.087	0.101	0.060	0.081	0.096	0.017
.	0.066	0.057	0.058	0.063	0.078	0.082	0.080	0.084	0.071	0.011
in	0.091	0.096	0.085	0.068	0.066	0.054	0.058	0.069	0.073	0.014
and	0.050	0.044	0.045	0.050	0.105	0.082	0.082	0.077	0.067	0.021
a	0.072	0.064	0.063	0.059	0.059	0.080	0.033	0.053	0.060	0.013

39

Burrows' Delta

The author's profiles in relative and their Z-score values.

	H	M
the	0.275	0.259
,	0.186	0.208
of	0.173	0.164
to	0.109	0.083
.	0.061	0.081
in	0.085	0.061
and	0.047	0.087
a	0.064	0.057

H	M
0.355	-0.308
-0.757	0.586
0.522	-0.379
0.769	-0.750
-0.955	0.934
0.819	-0.882
-0.927	0.945
0.319	-0.261

40

Burrows' Delta

The Delta distance for Doc #54

	H	M	D54	Δ(H)	Δ(D)
the	0.355	-0.308	0.573	0.218	0.881
,	-0.757	0.586	0.002	0.759	0.584
of	0.522	-0.379	-0.893	1.415	0.514
to	0.769	-0.750	-0.718	1.487	0.032
.	-0.955	0.934	0.769	1.724	0.165
in	0.819	-0.882	0.999	0.180	1.881
and	-0.927	0.945	-0.646	0.281	1.591
a	0.319	-0.261	-0.859	1.178	0.598
mean				**0.905**	**0.781**

41

Burrows' Delta

- Distance between two sub-corpora D (disputed) and D'
  If Δ is small, D and D' are written by the same author.

$$\Delta(D, D') = \frac{1}{m} \sum_{i}^{m} |Z(w_{iD}) - Z(w_{iD'})|$$

- Modification suggested (Hoover, 2004)
  - *m* must be greater than 150 (e.g., 800 – 4,000)
  - Ignore personal pronouns
  - Culling at 70% (words for which a single text supplies more than 70% of the occurrences).

Hoover, J. F. (2004). Delta Prime? *Literary and Linguistic Computing*, 19(4), 477-495.

42

42

Evidence #4: Delta

With 200 most frequent tokens

Rank	Distance	Author	Distance	Author
	L'amore molesto		*L'amica geniale*	
1	0.650	Starnone	0.524	Starnone
2	0.806	Brizzi	0.686	Veronesi
3	0.837	Milone	0.700	Balzano
4	0.850	Tamaro	0.721	Brizzi
5	0.874	Lagiola	0.726	Milone

43

Evidence #4

Delta model

- Applying Delta model based on the 200, 300, 400, or 500 most frequent tokens.
- Starnone is the author behind Ferrante's name.

- *L'amore molesto* (1992) is always assigned to Starnone.
- *Lacci* (2014) or *Scherzetto* (2016) is always assigned to Ferrante.
- Building 39 author profiles (wo Ferrante), all Ferrante's novels are assigned to Starnone.

Burrows, J. F. (2002). Delta: A Measure of Stylistic Difference and a Guide to Likely Authorship. *Literary and Linguistic Computing*, 17(3):267-287.

44

We found…

- Evidence #1
  Novel by novel
  Storia Della Bambina Perduta closed to *Lacci*
- Evidence #2
  By author: Ferrante closed to Starnone
- Evidence #3
  L'amica geniale closed to the vocabulary used more by Starnone
- Evidence #4 (Delta)
  Starnone

Domenico Starnone

with a prob. of 98%

45

Qualitative Analysis

Word	Corpus	Ferrante (6.5%)	Starnone (6.4%)
padre (father)	9,815	833 (8.5%)	1,170 (11.9%)
madre (mother)	8,246	1,104 (13.4%)	762 (9.2%)
perciò (therefore)	1,263	222 (17.6%)	254 (20.1%)
persino (even)	1,351	266 (19.7%)	205 (15.2%)
temere (fear)	1,345	274 (20.4%)	207 (15.4%)
tono (tone)	2,135	421 (19.7%)	286 (13.4%)
strunz	85	18 (21.2%)	63 (74.1%)

46

Is this correct?

- With different attribution methods and feature sets, Starnone always appears in the first (second) position in the ranked list…
- Assume that the accuracy of a method is p (e.g., p=80%) and the chance of an incorrect assignment is (1-p) (e.g., 20%).
- If the methods / feature sets are *independent*, probability that the r decisions are incorrect is $(1-p)\cdot(1-p)\ldots\cdot(1-p) = (1-p)^r$
- With p=80%, and r = 4, we have $0.2^4 = 0.0016$
  Prob. of correct = 1 - 0.0016 = 0.9984

47

Verification

- But is your conclusion falsifiable?
  Removing Starnone from the corpus…
- What do we expect?
- If we obtain always the same name in the top 3 positions… maybe we might have another possible author

48

Labbé's Distance wo Starnone (token)

Rank	Distance	ID	Name	ID	Name	
1	**0.103**	51	Ferrante	52	Ferrante	
2	**0.109**	50	Ferrante	51	Ferrante	
3	**0.114**	49	Ferrante	50	Ferrante	
4	**0.120**	50	Ferrante	52	Ferrante	
5	**0.127**	135	Veronesi	137	Veronesi	
6	**0.136**	43	Faletti	44	Faletti	
7	**0.137**	42	Faletti	44	Faletti	
8	**0.138**	49	Ferrante	51	Ferrante	
9	**0.142**	49	Ferrante	52	Ferrante	
10	**0.143**	42	Faletti	43	Faletti	
11	**0.145**	41	Faletti	42	Faletti	
…						
83	**0.210**	15	Benni	17	Benni	
84	**0.210**	56	Giordano	77	Milone	<-
…	**…**	…	…	…	…	
102	**0.216**	49	Ferrante	122	Sereni	<-

49

Verification: Delta

With 200 most frequent tokens, without Starnone

Rank	Distance	Author	Distance	Author
	L'amore molesto		*L'amica geniale*	
1	0.812	Tamaro	0.684	Veronesi
2	0.818	Brizzi	0.693	Balzano
3	0.843	Milone	0.714	Milone
4	0.861	Lagiola	0.722	Brizzi
5	0.880	Balzano	0.734	Nesi

50

Verification

- But is your conclusion falsifiable?
  Removing Starnone from the corpus…
- Applying different feature sets and attribution methods
  12 names appears at least once in the first rank
  Balzano, Mazzucci, Milone, Tamaro, Veronesi,
  Brizzi, Carofiglio, Giordano, Fois, Murgia, Raimo, Sereni

51

Epilogue

Elena Ferrante: 'I loved that boy to the point where I felt close to fainting'

In the first of a new weekly series, the novelist recalls her first love

by Elena Ferrante

Some time ago, I planned to describe my first times. I listed a certain number of them: the first time I saw the sea, the first time I flew in an aeroplane, the first time I got drunk, the first time I fell in love, the first time I made love. It was an exercise both arduous and pointless.

For that matter, how could it be otherwise? We always look at first times with excessive indulgence. Even if by their nature they're founded on inexperience, and so as a rule are not very successful, we recall them with sympathy, with regret. They're swallowed up by all the times that have followed, by their transformation into habit, and yet we attribute to them the power of the unrepeatable.

52

Who is the secret hand behind Elena Ferrante?

Domenico Starnone

(with a probability = 98%)

Savoy, J.: Is Starnone really the author behind Ferrante?. *Digital Scholarship in the Humanities*, 2018, 33(4), 902-918.

53