DRAWING ELENA FERRANTE'S PROFILE

Workshop Proceedings

Padova, 7 September 2017

Edited by
Arjuna Tuzzi and Michele A. Cortelazzo

PADOVA UP

PADOVA UNIVERSITY PRESS

Volume realizzato con il contributo del Dipartimento di Filosofia, Sociologia, Pedagogia e Psicologia Applicata dell'Università degli Studi di Padova.

Prima edizione anno 2018, Padova University Press

Titolo originale *Drawing Elena Ferrante's Profile.* Workshop Proceedings, Padova, 7 September 2017

Università degli Studi di Padova
via 8 Febbraio 2, Padova
www.padovauniversitypress.it

Progetto grafico
Padova University Press

ISBN 978-88-6938-130-0

Stampato per conto della casa editrice dell'Università di Padova – Padova University Press.

Drawing Elena Ferrante's Profile

Workshop Proceedings

Padova, 7 September 2017

Edited by

Arjuna Tuzzi and Michele A. Cortelazzo

PADOVA UP

Table of contents

Foreword

The case of the bestselling Italian author Elena Ferrante is a famous literary mystery much debated in many parts of the world. Since 2005 at least, various hypotheses have been advanced in the press about the real identity of Elena Ferrante, which is believed to be the pen name of an anonymous author. The subject has been discussed in major Italian newspapers like *La Stampa*, *L'Unità*, *Il Corriere della sera*, *Il Sole 24 ore*, *La Repubblica*, and in important foreign newspapers too, including *The Guardian* and *The New York Times*. The web has hundreds of thousands of pages about this case. Despite all the fuss in the media, the question of this author's identity has rarely been the object of scientific research, based on the methods for validating or refuting hypotheses typical of scientific research.

In 2016, an Italian research team embarked on a study suitable for submitting to the international scientific community for debate. It collected a corpus of 150 novels published in the last 30 years, written by 40 different Italian authors, and chosen according to precise parameters that took into account the main hypotheses emerging over the years concerning the real identity of Elena Ferrante, and the general scenario of contemporary Italian literature. To submit their findings to a broader scientific community for discussion, the authors adopted the well-established practice of presenting the results at specialist conferences and as peer-reviewed journal articles. They also went a step further: in the conviction that any worthwhile research is – by its very nature – transparent and available for debating, continuing, and confuting, as the case may be, they circulated their data to international experts of authorship attribution, profiling and analysis of textual data, inviting them to apply their own analytical methods to the material made available.

This volume is a collection of the contributions of various researchers who used various scientific methods to identify the author behind the novels by Elena Ferrante – a nom de plume that has become one of the most remarkable and often-discussed successes in the publishing world in recent years. The list of

the academics involved, in addition to the curators of this volume, Arjuna Tuzzi and Michele Cortelazzo (University of Padova), includes (in alphabetical order): Maciej Eder (Pedagogical University of Kraków – Polish Academy of Sciences, Poland), Patrick Juola (Duquesne University of Pittsburgh, PA USA), Vittorio Loreto and his research team, Margherita Lalli and Francesca Tria (University of Roma "La Sapienza", Italy), George Mikros (National and Kapodistrian University of Athens, Greece), Pierre Ratinaud (University of Toulouse II "Jean Jaurès" France), Jan Rybicki (Jagiellonian University of Kraków, Poland), and Jacques Savoy (University of Neuchâtel, Switzerland).

The results of the research conducted by this international group of experts were presented for the first time during the workshop *Drawing Elena Ferrante's profile*, held in Padua on 7 September 2017, as part of the 3rd IQLA-GIAT Summer School in *Quantitative Analysis of Textual Data*. The Summer School, directed by Arjuna Tuzzi and run by Padova University's *Dipartimento di Filosofia, Sociologia, Pedagogia e Psicologia Applicata* [Department of Philosophy, Sociology, Education and Applied Psychology], is an interdisciplinary program financed by the University of Padova. The exchange of ideas among the experts at the workshop, with the addition of contributions from 20 participants (from 11 different countries) attending the Summer School, provided the basis for the present publication.

Reading this volume, it is very interesting to see how the various contributions succeed in producing a genuinely interdisciplinary study on a concrete object of study. Not only were the authors of the contributions from all sorts of disciplines (linguists, social scientists, computer scientists, mathematicians, statisticians, physicists), they also conversed with one another from different analytical approaches. In addition, the vast majority of them do not speak Italian, so they worked on the corpus of novels completely blinded to the meaning of the words, trusting entirely to their methods for quantitatively analyzing textual data. Though they moved from different perspectives, their results supported and strengthened each other's like the different voices in a choir, leading to remarkably coherent and integrated conclusions.

Mirko Degli Esposti
University of Bologna "Alma Mater", Italy

It Takes Many Hands to Draw Elena Ferrante's Profile

Arjuna Tuzzi and Michele A. Cortelazzo
University of Padova, Italy

Abstract

Elena Ferrante is the pen name of a writer highly successful on the international stage, whose real identity has been kept secret by her publishers, E/O, for more than 25 years. For some time now, her fame and mystery have combined to attract the attention of readers, fans, critics, journalists, and academics all over the world, and various hypotheses have already been advanced as to who the hand behind this nom de plume might be. Elena Ferrante's work in general, and the question of her anonymity in particular, have been the object of only a limited number of scientific publications, however. Starting from a corpus of 150 novels by 40 different authors selected specifically for this research project, and from the results of previous qualitative and quantitative studies, this contribution contextualizes Elena Ferrante's work on the panorama of contemporary Italian literature. It also provides further elements to support the hypothesis that, based on what we know now, Domenico Starnone is the writer with the greatest affinity in style and content with the novels signed by Elena Ferrante. The present article also explains why our Italian research group shared our data and preliminary results with international experts on textual data analysis, authorship attribution, and profiling, who worked independently on the same data and added their own contribution to this case study.

The big issue of identity

It is intriguing to find that Elena Ferrante wrote the definition of *Identity* included among the "Authors' entries" in the 2016 edition of the Zingarelli dictionary, published in 2015 (Zingarelli, 2015, p. 1058):

> L'identità è la colla della molteplicità. Colla trasparente. Basta uno sguardo per accorgersi che negli occhi del singolo, dietro l'etichetta del nome e cognome, c'è una folla molto varia di spettri: io sono io è una semplificazione, utilissima per tenere in ordine noi stessi ma, come ogni possibile identità dentro cui ci ingabbiamo o siamo ingabbiati (sessuale, religiosa, nazionale, politica, sociale), è limitativa, ci impoverisce. Separarsi da io sono io almeno per un po', uscire da quel recinto specialmente nelle attività di invenzione o reinvenzione del mondo, apre uno spazio sterminato dove niente e nessuno resta identico e poi identico e poi ancora e sempre identico.
>
> [Identity is the glue of multiplicity. A transparent glue. A glance is enough to see in a person's eyes, behind the label of their name and surname, that there is a whole host of different ghosts. I am who I am is a simplification, very useful for keeping ourselves in order, but – like every possible identity in which we cage ourselves, or are caged (sexual, religious, national, political, social) – it is restrictive, we are impoverished by it. To move away from 'I am who I am', for a while at least, to escape that cage, especially in activities in which we invent or reinvent the world, opens up a limitless space where nothing and nobody remains the same, always the same, and the same again.]
>
> Elena Ferrante

That Zanichelli (the publishers of the Zingarelli dictionary) should choose Elena Ferrante to write this definition suggests that they have a good sense of humor. Or maybe they intended to be provocative, or they might have decided to entrust the task of defining *identity* to a real expert. Whatever their motives, the publishers requested a definition of *identity* from an author who uses a pen name. And her definition refers to the *multiplicity* of personalities that can hide behind a name and surname. This looks like a clue, one of many breadcrumbs left along a trail, suggesting without saying, arousing our curiosity, throwing us off track, and more or less deliberately nourishing the fuss around the famous name of Elena Ferrante.

Her fame stems largely from the success of her literary works, however. Her first two novels [with the titles of the English translations in brackets], *L'amore molesto* (Ferrante, 1992) [Troubling Love, 2006], and *I giorni dell'abbandono* (Ferrante, 2002) [The Days of Abandonment, 2005], were made into movies of the same name, directed by Mario Martone and Roberto Faenza, respectively. After her third novel, *La figlia oscura* (Ferrante, 2006) [The Lost Daughter, 2008], the author embarked on an extremely successful four-volume saga: *L'amica geniale. Infanzia, adolescenza* (Ferrante, 2011) [My Brilliant Friend, 2012]; *Storia del nuovo cognome. L'amica geniale volume secondo* (Ferrante, 2012) [The Story of a New Name, 2013]; *Storia di chi fugge e di chi resta. L'amica geniale volume terzo* (Ferrante, 2013) [Those Who Leave and Those Who Stay, 2014]; and *Storia della*

bambina perduta. L'amica geniale volume quarto (Ferrante, 2014) [The Story of the Lost Child, 2015). Episode after episode, this saga definitively crowned the author as an international star. The opening novel had been quietly submitted for consideration for the *Premio Strega* [Strega Prize] in 1992, while in 2015 the fourth volume of *L'amica geniale* (Ferrante, 2014) was shortlisted and classified in third place. In 2007, Tullio De Mauro included Ferrante's first novel *L'amore molesto* in his *Primo tesoro della lingua letteraria italiana del Novecento* [First Companion to Italian 20th-Century Literature], based specifically on the most significant competitors for the *Premio Strega*. In addition to her seven novels, Elena Ferrante has also published a children's story, *La spiaggia di notte* (Ferrante, 2007) [The Beach at Night, 2016], and a collection of meta-literary writings, mainly in the form of interviews, essays and letters, entitled *La Frantumaglia* (latest Italian edition: Ferrante, 2016) [Frantumaglia. A writer's journey, 2016].

The author became very popular in the United States, where it could be said that she was appreciated more and sooner than in her home country. According to many, this was also thanks to the quality of her translator, Ann Goldstein. In the US, and abroad generally, she is much loved and often quoted, even by exceptional people like the American ex-First Lady, Hillary Clinton. In 2014, the magazine *Foreign Policy* added the author's name to its list of leading global thinkers of the year.

Elena Ferrante's success in Italy has been reinforced by her enormous success abroad, and the mystery of her identity has certainly contributed to her popularity. With time, her publishers have established a real legend, giving rise to what some have called *Ferrante fever*. In 2017, this *Ferrante fever* was even converted from a brand into the title of a docufilm directed by Giacomo Durzi. A television series inspired by Ferrante's saga *L'amica geniale* has been announced too, directed by Saverio Costanzo (with the Neapolitan author Francesco Piccolo as one of the screenwriters). Since January 2018 Elena Ferrante has also been writing a weekly column in *The Guardian*.

The story and worldwide success of Elena Ferrante's books have become a topic that is not only of interest to her readers and fans, but also provides an opportunity to analyze the phenomenon as an object of scientific research. Elena Ferrante deserves to be studied using scientific criteria to fully understand the specificities of her work and the reasons for her success.

The author's real identity is kept jealously secret by her publishers, E/O, with a remarkable degree of discretion and determination that has persisted for more than 25 years. For some time now, the case of Elena Ferrante has attracted the attention of readers, fans, critics, journalists, and academics, and various hypotheses have been advanced in the Italian and foreign press as to her real

identity. These theories have naturally given rise to controversies, which have sometimes developed into full-blown confrontations between those who consider it worth trying to establish who hides behind the pen name of Elena Ferrante and those who believe we should respect the author's wish to remain anonymous and be "left in peace", as she clearly prefers to appear to the world only through her works. Both attitudes are based on well-founded arguments. On the one hand, there is the author's explicit intention to draw attention away from herself, almost obliging her readers to focus their interest on her texts alone. On the other, many critics are of the idea that a literary work should not be separated from its author: although the text can undoubtedly have its intrinsic value, it is difficult for the literary critic to consider it as something wholly separate from its writer.

Over time, various people have been suspected of being behind Elena Ferrante's books, including: novelists like Guido Ceronetti, Erri De Luca, Francesco Piccolo, Michele Prisco, Fabrizia Ramondino, and Domenico Starnone; essayists and academics like Goffredo Fofi, Marcello Frixione, and Marcella Marmo; screenwriters and directors like Mario Martone and Linda Ferri; journalists like Daria Bignardi; translators like Anita Raja; and even the publishers themselves, Sandra Ozzola and Sandro Ferri (to name just a few of the more or less credible names circulating in the press and on the web). There have also been numerous "investigators" (journalists and academics) who have worked on the Elena Ferrante case, using various methods and with more or less variable results. Some of the suggestions, documentation, data, and comparisons announced and discussed in newspaper articles and blogs have been based on accurate analyses and methods, that make them worthy of a scientific debate.

A first, amply-documented study was published in an article in the Italian newspaper *La Stampa* (Galella, 2005). Galella compared some pages of *L'amore molesto* by Elena Ferrante (1992) and *Via Gemito* by Domenico Starnone (2000). Since marked thematic and lexical similarities emerged from this comparison Galella concluded that this latter author had written the novels under the pen name of Elena Ferrante. A year later Galella, in an article published in the Italian newspaper *L'Unità*, reported the results of a quantitative study conducted by the physicist Vittorio Loreto. Using compression algorithms, Loreto tested the similarities between the novels by Elena Ferrante and those of Domenico Starnone, Goffredo Fofi, Fabrizia Ramondino, Michele Prisco, and Erri De Luca. The outcome was a strong similarity between the writings of Ferrante and Starnone, which were clearly distinguishable from those of the other authors, and resembled each other enough to appear like the work of the same author. Similar conclusions were reached in 2016 in another, more limited study conducted by the Swiss company *OrphAnalytics* (2016), which published a report on the topic on its website.

Through further examples of thematic and lexical similarities, Galella's qualitative perspective has been confirmed by Simone Gatto (2006; 2016) in two essays published in the blog *Lo specchio di carta*, which is the expression of the *Osservatorio sul romanzo italiano contemporaneo* [Observatory on the Contemporary Italian Novel] at the University of Palermo.

A different hypothesis has been proposed by Marco Santagata, professor of Italian Literature, in an article that appeared in *La Lettura* literary magazine of *Il Corriere della Sera*. Santagata (2016) retrieved from the second volume of the saga *My brilliant friend* some relevant details about Elena's experience in Pisa as a student of the Scuola Normale Superiore. Santagata concluded that the author was probably Neapolitan, woman and had attended the Scuola Normale in Pisa during the 1960s, but before 1966. Santagata's identikit coincided with Marcella Marmo, who had since become professor of contemporary history at the Federico II University in Naples.

Then came the journalist Claudio Gatti (2016), who focused on extratextual data (assets and liabilities), publishing the results of his investigation in *Il Sole 24 ore* (and some foreign papers), which indicated that the payments made by the publishers E/O to Anita Raja (a translator from German for this same publishing house and also the wife of Domenico Starnone) could only be explained by identifying her with Elena Ferrante.

So the much-discussed mystery surrounding the persona behind Elena Ferrante has been the object of numerous investigations. Almost all of the debate regarding her identity has been conducted in the world of journalism, while the question has not been the object of extensive scientifically-designed research, including an open discussion among peers at conferences and their publication in scientific journals. At the time of the *Conference* of the *International Quantitative Linguistics Association* (IQLA) held in Trier in 2016 (QUALICO2016), our research group presented the case of Elena Ferrante on the strength of a large corpus of Italian contemporary literature (Cortelazzo et al., 2018). Drawing on the interdisciplinary nature of the research team, we were able to work with both quantitative and qualitative methods (Cortelazzo and Tuzzi, 2017; Tuzzi and Cortelazzo, 2018; Cortelazzo, Mikros and Tuzzi, 2018).

Why study Elena Ferrante?

The case of Elena Ferrante is interesting from both the stylistic and the stylometric standpoints. From the stylistic standpoint, we might wonder whether the international success of Elena Ferrante's novels (which far exceeds that of other products of Italian contemporary literature) stems simply from the fascination of Naples and its social dynamics in the 1950s, or whether the reason

lies in the author's personal writing style. From the stylometric standpoint, it seems worth studying and measuring the similarities and differences between the linguistic features of Elena Ferrante's works and those of other contemporary Italian novelists.

It has to be said that, as an unavoidable side effect of seeking such affinities with other texts, elements will emerge that may help to identify the mysterious author behind this nom de plume. Research, by its very nature, is always attracted to mysteries, and constantly engaging in formulating new problems as well as seeking new solutions for old problems. But researchers take a special interest in situations in which they can put the tools at their disposal to the test, and contribute to their improvement. This is certainly the case of Elena Ferrante's appeal: our interest in the mystery of her identity was not prompted by mere morbid curiosity, but because it gave us the chance to test the quantitative and qualitative methods available for accurately identifying common and distinctive features in a set of texts collected into a corpus.

The issue of Elena Ferrante's identity is highly complex, if we think that the people suspected of being behind her name have included not just novelists, but also essayists, translators, critics, journalists, film directors, and academics. The enigma thus obliges scholars to extend the analysis to fields hitherto scarcely visited by anyone involved in authorship attribution. But, apart from the authorship attribution problem in a strict sense, other aspects mentioned by the literary critics who have analyzed Ferrante's works need to be tested for profiling purposes, i.e. the author's gender, age, and geographical origins. The question of gender is particularly engaging because many critics (and many readers) see the novels as representing the protagonists' relationships with family and friends from a strictly feminine perspective (Ceccoli, 2017; Chemotti, 2009; Dow, 2016; Lee, 2016). It would be remarkable if it were to emerge that the author (or even co-author) of the books by Elena Ferrante is actually male. Another point on which the critics agree concerns the appeal of Naples as a backdrop to Ferrante's stories and the resulting atmosphere (Alfonzetti, 2018; Benedetti, 2012; Caldwell, 2012; Cavanaugh, 2016; Falotico, 2015; Librandi *in press*; Ricciotti, 2016). It is well worth further investigating whether her environment places Elena Ferrante, without particular distinction, among other Neapolitan authors, or writers who set their stories in Naples, whether it associates her with one other such author in particular, or whether it is an attribute all her own.

In short, we felt duty bound to apply a scientific method to studying such a remarkable literary phenomenon because, in the 26 years since the publication of her first volume, the scientific output on the works of Elena Ferrante is surprisingly limited; though it has been on the rise in recent years (see Bullaro and Love, 2016). To neglect Elena Ferrante as an object worthy of scientific (and ac-

ademic) research seems far worse than to risk violating the author's (presumed) wish to remain anonymous.

The use of a mixed method

It seemed to us that the best way to approach our task would be to conduct quantitative analyses. The world of the digital humanities and digital methods is constantly expanding and has contributed a great deal to the diffusion of new research methods. In the sphere of the humanities, however, the treatment of texts (and literary texts, in particular) using quantitative methods still raises doubts and objections. The promoters of quantitative approaches are accused of reducing the complexity and beauty of literary texts with their arid mathematical formulas. The process is often considered pointless too, because analyzing a literary text in depth simply demands the sensitivity and the intuition of an expert reader, and preferably of a literary critic with years of training on great variety of texts. An expert's intuition can sometimes be misleading, however (and quantitative methods can sometimes fail too).

In our opinion, the best way to proceed is to combine quantitative with qualitative observation. If we are expert readers we can identify the style, and pinpoint the idiosyncrasies and habits of our favorite authors, but the quantitative analysis of textual data enables us to test these intuitions systematically and on a vast scale (useful for the purposes of a confirmatory analysis). Those who use quantitative methods need a dose of intuition too because we cannot compare texts chosen at will, or at random: we need a priori hypotheses to guide the construction of the corpus, and these hypotheses are necessarily of qualitative type. In addition, not everyone has the expert reader's intuitive skills, and quantitative methods (and the corresponding software packages) can supplement our intuition and point our qualitative research (seen from the perspective of an exploratory investigation) in unexpected and unexplored directions. Quantitative analysis can also offer a reliability that qualitative analysis could never match. For instance, it is thanks to the use of quantitative text analysis methods that a researcher can say that a phenomenon (such as a particularly significant word) is present or absent in a text on the grounds of a systematic observation uninfluenced by any degree of discretionality or subjectivity, or failings of human memory.

Briefly, given the opportunity to treat texts as data, we can:

1) extend the corpus forming the object of the study well beyond the dimensional boundaries manageable using conventional qualitative analytical methods; and

2) focus on problems that can only be studied on large corpora.

The corpus

For the present study on Elena Ferrante's novels on the landscape of contemporary Italian literature, a corpus was collected ad hoc comprising novels originally written in the Italian language (i.e. no translations from other languages were considered) over a period of 30 years, from 1987 to 2016 (with only four exceptions, which were added to reinforce the subcorpora of three authors: Michele Prisco, Dacia Maraini, and Marta Morazzoni). During the novel selection phase, we opted to include in our corpus:

- novels written by authors presumably from the same geographical region (Naples and the surrounding area);
- novels written by women;
- blockbusters (best-sellers, award-winning novels);
- novels written by authors praised by the literary critics; and
- novels written by writers suspected of being Elena Ferrante.

This large corpus was therefore based on accurately-chosen criteria and has characteristics that should be adequate for the purpose of placing Elena Ferrante in the context of the best-known Italian contemporary authors, and containing elements for comparing Elena Ferrante with twelve other female authors, and with ten other Neapolitan authors.

The corpus includes 150 novels written by 40 different authors: in addition to Elena Ferrante, Eraldo Affinati, Niccolò Ammaniti, Andrea Bajani, Marco Balzano, Alessandro Baricco, Stefano Benni, Enrico Brizzi, Gianrico Carofiglio, Mauro Covacich, Erri De Luca, Diego De Silva, Giorgio Faletti, Marcello Fois, Paolo Giordano, Nicola Lagioia, Dacia Maraini, Margareth Mazzantini, Melania Mazzucco, Rossella Milone, Giuseppe Montesano, Marta Morazzoni, Michela Murgia, Edoardo Nesi, Paolo Nori, Valeria Parrella, Francesco Piccolo, Tommaso Pincio, Michele Prisco, Christian Raimo, Fabrizia Ramondino, Ermanno Rea, Tiziano Scarpa, Clara Sereni, Domenico Starnone, Susanna Tamaro, Chiara Valerio, Giorgio Vasta, Sandro Veronesi, and Simona Vinci. It is a large corpus in size, including nearly 10 million word-tokens (9,837,851), so the novels in the collection contain an average of 65,586 words each. The longest novel consists of nearly 200 thousand word-tokens (*Io uccido* by Giorgio Faletti) and, with one exception (*Behave* by Valeria Parrella), all the novels contain more than 10,000 word-tokens. The corpus expresses an overall vocabulary of 159,149 different word-types obtained using the tokenizing, and uppercase letter reducing system in the *Taltac* software package (Bolasco, 2010). The type-token ratio of the corpus as a whole amounts to 1.6%, which corresponds to a mean overall frequency of approximately 62 occurrences of each word-type. The words that occur only once (hapax legomena) amount to 58,723, corresponding to 37% of the vocabulary.

It is important to make it very clear that, for the purposes of authorship attribution, our corpus can only be useful if the author behind the nom de plume of Elena Ferrante is a writer who has also published other significant narrative works during the same period, that have been included in our corpus. This is plausible, but by no means the only possible hypothesis. For instance, our mystery author may have only published the Ferrante novels, in which case we would lack the elements needed for a comparison, and it would be impossible to establish the author's identity by comparing a set of texts. Or the author of the Ferrante novels may have published other types of text (not novels) under his/her real name, but then we would need to draw comparisons with a completely different corpus from the one used in the present study. For this reason, we are now working on a (more circumscribed) corpus enabling us to compare the texts of "suspects" like Anita Raja or Marcella Marmo with the meta-literary texts of the *Frantumaglia* (Cortelazzo, Mikros and Tuzzi, 2018).

Results of the research conducted by the group at the University of Padova

Our study was conducted on the above-described corpus using mainly quantitative methods, with which we associated some qualitative considerations. At a first stage it had two aims:

1. to place Elena Ferrante's novels on the landscape of contemporary Italian literature, within the setting of a large collection of 150 novels by 40 different authors;
2. to study Elena Ferrante in relation to this body of literature by gender and region, as we expected diastratic (gender) and diatopic (region) variations to play an important part in terms of topics and linguistic features.

It was only at a second stage that, in the light of the results obtained, the study turned to with the question of authorship attribution. The main purpose of this step was to test the hypotheses already advanced in journals and blogs concerning the author's identification. Our results are briefly outlined, listing the methods adopted, and the results obtained in each case (Tuzzi and Cortelazzo, 2018).

Content mapping based on correspondence analysis

As in other, previous works of ours, we first used an exploratory data analysis (EDA) to compare authors and novels. This involves content mapping based on correspondence analysis (CA), a well-known multivariate statistical tech-

nique that uses the occurrences of words in the texts and displays a two-way contingency table by means of coordinates on Cartesian axes that locate the rows (words), and columns (authors or novels) on a plane

The results of the CA applied to the whole corpus reveal three authors who appear to be the most original: Paolo Nori, Giorgio Faletti, and Elena Ferrante. In particular, Ferrante shows the greatest individuality in her themes and writing. We can find some Neapolitan (or more generally southern Italian) writers close to Elena Ferrante: Prisco, Rea, Carofiglio, Starnone, and also De Luca, Montesano, Parrella, plus the Milanese writer Balzano, whose family comes from the south. Among the significant authors who are not southern Italians, we find Clara Sereni. The novelist coming closest to Elena Ferrante, and consequently the most similar in terms of their lexical profiles, is Domenico Starnone.

When our analysis was applied to the sub-corpus of 13 female writers, the position occupied by Elena Ferrante emerges very clearly: she stands in splendid isolation, very obviously unlike any of the other female writers, apart from a very slight resemblance with Rossella Milone (a young author who may have taken Elena Ferrante as a model). When single novels were considered, a moderate affinity emerged only with certain novels by Michela Murgia (*Chirù*, 2015), and Clara Sereni (*Via Ripetta*, 2015).

On the other hand, if we look at the geographical setting and compare Elena Ferrante's novels with those of another ten authors from the Naples area and their novels, we again find her isolated from the rest of the group and close to Starnone. Most of her works are again very close to one another. Only her first novel (Ferrante, 1992) seems to stand a little way from the others, in a position close to novels by De Luca, Milone, and De Silva, though lying not so far from Ferrante's other novels. We can also mention a resemblance with the novels published by Domenico Starnone after 1993. A particular feature of Starnone's works lies in that they take two clearly distinct positions: his earlier works (1987-1991) are situated in a separate group, while those written since 1993 are located close to Elena Ferrante's novels.

Text clustering and inter-textual distance

Based on some prior experiences (Tuzzi, 2010), we exploit Labbé's intertextual distance (Labbé and Labbé, 2001) in the iterative version previously proposed by our group (Cortelazzo et al., 2013). We worked with repeated measures on numerous samples of equal-sized chunks. This procedure involves extracting a sample of 150 text-chunks (one for each novel) of the same size for each replication and calculating a distance for each pair of text-chunks.

This calculation on samples is repeated numerous times, and the distance between each pair of novels is obtained from the mean of all the distances calcu-

lated between pairs of their text-chunks. At the end of this iterative procedure we obtain a square matrix that includes 150 x 150 cells and reports distances between each pair of novels. This matrix can be used to classify the texts automatically, by means of an agglomerative hierarchical cluster algorithm with complete linkage. In our case, the algorithm identified a cluster that includes all the novels written by Ferrante and many of those written by Starnone, except for his first three works, which form a small separate cluster. The results of the CA were therefore confirmed.

It is useful to try reading the matrix of the distances from another perspective, i.e. using a ranking system: the rows (or columns) of the square matrix represent the distances between one novel and all the others so, for a given novel, all the others can be sorted from the closest (minimum distance) to the farthest away (maximum distance). Looking at the first positions, we can see that all except the first of Elena Ferrante's works mainly resemble another work by the same author. In particular, all four novels in her saga are very similar to one another (as might be expected in a story in four episodes), while positions coming next show the greatest resemblance with Domenico Starnone's latest novels. Ferrante's first novel, *L'amore molesto*, has a rather particular profile, appearing more similar not to one of her own novels, but to *Eccesso di zelo*, a novel written by Domenico Starnone and published in 1993 (which also came second for Ferrante's next two novels, *I giorni dell'abbandono* and *La figlia oscura*). The rankings relating to Domenico Starnone novels from *Eccesso di zelo* onwards (Starnone, 1993) showed at least one work by Elena Ferrante in first place, as if to suggest that Elena Ferrante resembles Domenico Starnone even more than Starnone himself. The rankings of the earlier works by Starnone show a very different picture: his first work (*Ex cattedra*) bears no resemblance to Elena Ferrante's novels, while her works begin to appear in the rankings for Starnone's second and third novels, but only in seventh or eighth place. The picture becomes even clearer if we measure the inter-textual distance based on the grammatical words only, after excluding all the words that convey content (names, adjectives, verbs, adverbs). This goes to show that the similarities between the novels written by Ferrante and Starnone are independent of any resemblances in their setting or how the story develops.

Tests on non-literary texts

Our recent research is now continuing in new directions. Based on a new, more focused and circumscribed corpus (Cortelazzo, Mikros and Tuzzi, 2018), we are now considering a set of candidates who are not strictly novelists, starting with Anita Raja, Marcella Marmo and the editors at the E/O publishing house. This new corpus has been used to compare Elena Ferrante's collection of

meta-literary fragments in *La frantumaglia* (Ferrante, 2016) with a set of essays, newspaper articles, and interviews by other writers. We used a profiling technique based on machine learning (ML) and support vector machine (SVM). Our preliminary results suggest that more than one author might have contributed to Ferrante's collection, seemingly at least one man and one woman, and the author(s) would appear to come from Naples, and be over 60 years old. Looking again at the main authors suspected of being Elena Ferrante, results point to Domenico Starnone and Anita Raja, while it seems that we can rule out Marcella Marmo. The work of the editors at E/O is relevant too, as they may have had a role in preparing the texts in *La Frantumaglia*.

Moving from a quantitative to a qualitative analysis: lexical affinities

Having completed our numerous quantitative analyses, we returned to conducting more conventional qualitative investigations, especially with a view to describing Elena Ferrante's lexical characteristics and affinities with other contemporary authors (Cortelazzo and Tuzzi, 2017).

The availability of a large corpus enabled us to confirm the lexical affinities identified by Luigi Galella (2005; 2006), and Simone Gatto (2006; 2016), but also to see whether these words were exclusive to Ferrante and Starnone, or shared with other authors too.

We thus confirmed the following previously-identified lexical affinities [in brackets a literary translation is reported]:

- the syntagms *collo filettato* [threaded neck] and *foglio di compensato* [plywood], contained in *L'amore molesto* by Ferrante and *Via Gemito* by Starnone (and only in these two novels), in descriptions of the objects used by the protagonists' father when he painted;
- the syntagm *vestaglia verde* [green housedress] used by Ferrante in *L'amore molesto*, *L'amica geniale*, and *Storia della bambina perduta*, and by Starnone in *Eccesso di zelo,* and *Via Gemito*; the mentions of this housedress in *Storia della bambina perduta* and *Eccesso di zelo* also share the adjective *vecchia* [old]: *vecchia vestaglia verde*;
- the word *argentiera* [silverware cabinet], which appears in Elena Ferrante's *L'amore molesto*, *Storia del nuovo cognome*, and *Storia della bambina perduta*, and in *Via Gemito* (four times) and *Scherzetto* by Domenico Starnone. While it is true that, where these two authors come from, this piece of furniture was once commonplace and generally known by the name of *argentiera*, as Starnone declared in an interview (Baudino, 2005, p. 27):

> Io e la Ferrante veniamo dalla stessa area lessicale, siamo napoletani. Il mobile che entrambi chiamiamo "argentiera" era diffusissimo in tutte le famiglie, e portava quel nome anche se magari di argenti non ne aveva ospitati mai
> [Ferrante and I come from the same lexical area, we are both from Naples. The piece of furniture that we both call "argentiera" was very common in every family, and went by that name even if it may never have contained any silverware.]

This objection becomes less convincing when we find that none of the other Neapolitan authors in our corpus had ever happened to name this particular piece of furniture.

There were also numerous other words or expressions in our corpus shared exclusively by the two authors in question, however (or with only sporadic occurrences elsewhere). To mention just a few, we found: the nouns *buffé* [sideboard], *calettatura* [splicing], *càntaro* [pitcher], *giravite* [screwdriver], *feticismo* [fetishism], *frantumaglia* [fragments], *latrocinio* [theft], *malodore* [stink], *mamozio* [dolt], *psicologismo* [psychologism], *risatella* [little laugh], *santodio* [good God!], and *turpitudine* [turpitude] (but also *album di fotografie* [photo album], *asso pigliatutto* [winning ace], and *mattonelle sconnesse* [uneven tiles]); the numeral *centoquarantotto* [a hundred and forty-eight]; the adjectives *apprezzatissimo* [much appreciated], *fonatorio* [phonatory], *gialloverdognolo* [greenish yellow], *sfottente* [mocking], *taglientissimo* [very sharp], *valutativo* [appraising] (with reference to a *sguardo* [gaze]), and *vagolante* [splicing]; the verbs *bamboleggiare* [act like a doll], *femminilizzare* [feminize], *riplasmare* [reshape], *sbruffoneggiare* [brag], and *spetazzare* [fart]; the adverbs: *buffamente* [funnily], *calcolatamente* [calculatedly], *caramente* [dearly], *contraddittoriamente* [contradictorily], *fievolmente* [weakly], *lietamente* [cheerfully], *meditatamente* [painstakingly], *saviamente* [wisely], *sforzatamente* [effortfully], and *soffertamente* [with difficulty]; the expressions: *aaaah* (with four "a"), *ciuciù* [choo choo], and *tottò sulle manine* [a gentle rap over the knuckles].

It is worth adding a comment on some of these terms. First of all, *risatella* [little laugh, noun] is a a characteristic word of southern Italy, and the Naples area particularly (as confirmed by its inclusion in vocabularies of Neapolitan dialect, e.g. Andreoli, 1887). It occurs in the corpus, in both the singular and the plural, but only in Ferrante (20 times) and Starnone (10 times). To find other occurrences in the literature, we need to move away from the novels selected for our corpus (e.g. it appears in *Passaggio in ombra* by Maria Teresa di Lascia, 1995), and also from the chronological period considered (we can find it in *Horcynus Orca* by Stefano D'Arrigo, published in 1975).

A similar line of reasoning can be adopted for *malodore* [stink], which is contained in our corpus – with this spelling – but only in Ferrante (12 times) and Starnone (5 times); elsewhere and not included in our corpus, we can find

it in Stefano D'Arrigo, once again, and in Gesualdo Bufalino (*Diceria dell'untore*, 1981). This is an even more distinctive lexical Neapolitanism, occurring in the variant *maleodore* in just one other Neapolitan author in our corpus, Francesco Piccolo (13 times).

The spelling is of interest in the case of *santodio* [good God!] too, which is a graphical variant of *santo dio* and a lexical variant of *santiddio* and *santo iddio* (an expletive used to express disappointment, indignation, impatience or surprise). We find no trace of *santiddio* (the form preferred by Umberto Eco, among others) in Ferrante-Starnone's works, nor of *santo iddio*. Instead, in Starnone we find *santo dio* (which occurs only once in Ferrante, but is also to be found in many other authors' novels) and also *santodio*, as one word (12 times). This latter form, *santodio* emerged in our corpus only one other time, in *I giorni dell'abbandono* by Elena Ferrante.

Another important case concerns the neologism *sbruffoneggiare*, meaning to brag, which was included in the Zingarelli dictionary for the first time in its 2012 edition, where its origin was dated back to 1992 – when the word appeared in Elena Ferrante's *L'amore molesto*, so she was probably the first to use this term. We find it again in *Storia di chi fugge e di chi resta*, the third volume of her saga, and also in two novels by Starnone, *Eccesso di zelo* and *Spavento* (not included in our corpus). In the rest of our corpus there was no sign of *sbruffoneggiare*. Generally speaking, we could say that this word is not only relatively new, but also rarely used. If we google *sbruffoneggiare* we find just 1,470 results, and if we look for the third person singular of the present tense, *sbruffoneggia*, we arrive at 3,210 results, which include the previous ones (search conducted on 9 March 2018).

There are some particular sequences that become significant too. To give a couple of examples, one is *tra la mandibola e la clavicola* [between the jaw and the collarbone], to indicate a part of the body in *L'amica geniale* (*Appena aprì la finestra gli arrivò in faccia uno sbuffo di pioggia e sul lato destro del collo, proprio a mezza strada tra la mandibola e la clavicola* [As soon as he opened the window a gust of rain struck his face and someone plunged a knife into the right side of his neck, halfway between the jaw and the collarbone], and also in *Denti* by Domenico Starnone (*Le gettò il braccio libero sulle spalle e si tuffò d'impeto col naso e la bocca tra la mandibola e la clavicola di lei* [He threw his free arm over her shoulders and dived with his nose and mouth straight between her jaw and her collarbone]). This is such an unusual way of indicating this part of the body that if we Google it we will only find 10 results (including those of the two authors in our corpus). Another example is *di scempio e di sangue* [of massacre and blood], which is so unusual that, if we Google it, in the immense body of texts available in Internet we can find only the examples contained in the works by Ferrante

(*Storia del nuovo cognome*) and Starnone (*Denti*). In both cases, the phrase is associated with the word *immagini* [images] (respectively, "*immagini improvvise di scempio e di sangue, molto presenti nei suoi quaderni*" [sudden images of massacre and blood, which were very frequent in her notebooks], and "*avevo cominciato a coltivare immagini di scempio e di sangue già verso gli otto anni*" [I had already begun to dream up images of massacre and blood by the time I was turning eight years old]).

Open questions

Analyzing these 150 novels collected in our corpus generated some interesting results. In all of our analyses, Elena Ferrante shows traits of originality in both style and content. We lack the elements needed to say for sure, but this is plausibly one of the main reasons for her success.

Our analyses also confirmed the remarkable affinities between the works of Elena Ferrante and those of Domenico Starnone. All the measures we used to test the similarities between their novels indicated that they are almost inextricably entwined. Our data thus provide more systematic arguments to support the claim advanced already in 2005 in the light of quantitative and qualitative studies: the distant reading used in our study supports the close reading that led Luigi Galella and Simone Gatto to identify detailed thematic, contextual and lexical affinities between the works of the two authors. Our research effort thus satisfies a fundamental principle of scientific research, that of testing and validating the results of previous studies by the replicating experiments, using the same or different methods (as in our case).

The similarity between the two authors is reinforced by the fact that Domenico Starnone's writing style has changed considerably since Elena Ferrante's books came on the scene. Such a difference between Starnone's earlier and later literary production can only have been partly due to his first works being collections of short stories about school life (many of which were initially published in newspapers and magazines). In fact, one of his early works analyzed in our corpus was Starnone's first novel, *Il salto con le aste*, which seems entirely consistent with his production written before 1992. Moreover, when we do not consider content words the degree of similarity increases and appears earlier.

Our more recent study also supports Claudio Gatti's conclusions that Anita Raja might have a role in this story. So, Elena Ferrante's books may be the work not of a single author, but of some form of cooperation (though what form this may take is not easy to imagine) involving at least two authors, one (Anita Raja) identified by Gatti's investigation into the accounts of Elena Ferrante's publishers, another by stylometric studies (Domenico Starnone).

Like any other research project, our work also raises a number of questions. When compared with a large corpus of contemporary Italian novelists, Elena Ferrante shows some remarkably individual traits. What makes her works so original? Who writes her books, and how are they produced? What is Elena Ferrante? When compared to other women, she seems distinct from the others. Is Elena Ferrante really a woman? When compared with other Neapolitan writers, only Starnone bears some resemblance to her. Is the regional setting of her books relevant? The remarkable similarities between Starnone and Ferrante can also be placed on a common time scale, starting in the early 1990s. What happened around 1990? Come to that, who is Domenico Starnone, and how do his books come into being?

Passing the baton to an international research group

As we believe in open science and practices designed to make processes and results transparent and accessible to investigators outside our research team, we circulated our results to some of the most outstanding international scholars of textual data analysis, author attribution, and profiling, asking these experts to apply their own analytical methods to our dataset. Our research group was thus joined by seven other researchers from six countries, each with their own expertise and different backgrounds: Maciej Eder and Jan Rybicki (Poland), Patrick Juola (United States), Vittorio Loreto (Italy), George Mikros (Greece), Pierre Ratinaud (France), and Jacques Savoy (Switzerland). We met in Padua at the IQLA-GIAT Summer School in *Quantitative Analysis of Textual Data* (3rd edition), a project funded by the University of Padova, where we had the chance to compare the results we had achieved using our various different methods. The continuation of the work and the rest of this book is entrusted to their experience and knowledge.

The workshop *Drawing Elena Ferrante's Profile* attracted some attention from the national and international press. The national daily *La Repubblica* dedicated a page (De Santis, 2017) to the workshop's findings, while *Il Mattino di Padova* published a résumé written by Michele Cortelazzo (Cortelazzo, 2017). The workshop was not mentioned by *Il Corriere della sera* (which had committed to sustaining Marco Santagata's hypothesis in 2016), *Il Sole 24 ore*, the daily in which Claudio Gatti had published his investigation, or *La Stampa*, although our quantitative analysis confirmed what Luigi Galella had written in the same paper already in 2005. Elsewhere, our workshop was also a news item in the Greek *Ta Nea* and *To Vima* papers, and in the Swiss *L'Express*. The conclusions reached at the meeting in Padua were also mentioned on several occasions by other media in various countries. For instance, the Swedish national broadcast-

ing company (SVT) made an announcement (subsequently echoed by several Swedish newspapers), the Italian national broadcasting company (RAI), on its Rai News website, and the online daily *Il Post*, as well as Danish (*Politiken*) and Dutch (*de Volkskrant*) newspapers.

Now it is time for the debate to move from the pages of newspapers and blogs to the world of scientific publications. The contributes of the present volume aim to offer a broad, detailed, interdisciplinary, and international contribution in this sense.

List of novels

Eraldo Affinati, *Campo del sangue* (1997), *L'uomo del futuro* (2016); Niccolò Ammaniti, *Fango* (1999), *Ti prendo e ti porto via* (1999), *Io non ho paura* (2001), *Come Dio comanda* (2006); Andrea Bajani, *Se consideri le colpe* (2007), *Ogni promessa* (2010), *Mi riconosci* (2013); Marco Balzano, *Il figlio del figlio* (2010), *L'ultimo arrivato* (2014); Alessandro Baricco, *Castelli di rabbia* (1991), *Oceano mare* (1993), *City* (1999), *Questa storia* (2005); Stefano Benni, *Il bar sotto il mare* (1987), *Margherita Dolcevita* (2005), *Di tutte le ricchezze* (2012); Enrico Brizzi, *Jack Frusciante è uscito dal gruppo* (1994), *L'inattesa piega degli eventi* (2008), *Il matrimonio di mio fratello* (2015); Gianrico Carofiglio, *Testimone inconsapevole* (2002), *Ad occhi chiusi* (2003), *Il passato è una terra straniera* (2004), *Ragionevoli dubbi* (2006), *Le perfezioni provvisorie* (2010), *Il silenzio dell'onda* (2011), *Il bordo vertiginoso delle cose* (2013), *Una mutevole verità* (2014), *La regola dell'equilibrio* (2014); Mauro Covacich, *A perdifiato* (2005), *Prima di sparire* (2008); Erri De Luca, *Tu, mio* (1998), *Tre cavalli* (1999), *Il giorno prima della felicità* (2009), *I pesci non chiudono gli occhi* (2011); Diego De Silva, *La donna di scorta* (1999), *Certi bambini* (2001), *Non avevo capito niente* (2007), *Mia suocera beve* (2010), *Sono contrario alle emozioni* (2011); Giorgio Faletti, *Io uccido* (2002), *Niente di vero tranne gli occhi* (2004), *Fuori da un evidente destino* (2006), *Io sono Dio* (2009), *Tre atti e due tempi* (2011); Elena Ferrante, *L'amore molesto* (1992), *I giorni dell'abbandono* (2002), *La figlia oscura* (2006), *L'amica geniale. Infanzia, adolescenza* (2011), *Storia del nuovo cognome. L'amica geniale volume secondo* (2012), *Storia di chi fugge e di chi resta. L'amica geniale volume terzo* (2013), *Storia della bambina perduta. L'amica geniale volume quarto* (2014); Marcello Fois, *Stirpe* (2009), *Nel tempo di mezzo* (2012), *Ex voto* (2015); Paolo Giordano, *La solitudine dei numeri primi* (2008), *Il corpo umano* (2012), *Il nero e l'argento* (2014); Nicola Lagioia, *Tre sistemi per sbarazzarsi di Tolstoj* (2001), *Riportando tutto a casa* (2009), *La ferocia* (2014); Dacia Maraini, *Memorie di una ladra* (1972), *La lunga vita di Marianna Ucria* (1990), *Buio* (1999), *Il treno dell'ultima notte* (2008), *La grande festa* (2011); Margareth Mazzantini, *Non ti muovere* (2002), *Venuto al mondo* (2008), *Mare al*

mattino (2011), *Nessuno si salva da solo* (2011); Melania G. Mazzucco, *Il bacio della Medusa* (1996), *Vita* (2003), *Un giorno perfetto* (2005), *Un giorno da cani* (2007), *La lunga attesa dell'angelo* (2008); Rossella Milone, *Poche parole, moltissime cose* (2013), *Il silenzio del lottatore* (2015); Giuseppe Montesano, *Nel corpo di Napoli* (1999), *Di questa vita menzognera* (2003); Marta Morazzoni, *La ragazza col turbante* (1986), *Il caso Courrier* (2005); Michela Murgia, *Il mondo deve sapere. Romanzo tragicomico di una telefonista precaria* (2006), *Viaggio in Sardegna. Undici percorsi nell'isola che non si vede* (2008), *Accabadora* (2009), *Ave Mary. E la Chiesa inventò la donna* (2011), *Chirù* (2015); Edoardo Nesi, *Rebecca* (1999), *Storia della mia gente. La rabbia e l'amore della mia vita di industriale di provincia* (2010), *L'estate infinita* (2015); Paolo Nori, *Bassotuba non c'è* (1999), *La matematica è scolpita nel granito* (2011), *Tredici favole belle e una brutta* (2012); Valeria Parrella, *Per grazia ricevuta* (2005), *Behave* (2011); Francesco Piccolo, *Storie di primogeniti e figli unici* (1996), *Allegro occidentale* (2003), *L'Italia spensierata* (2007), *La separazione del maschio* (2008), *Momenti di trascurabile felicità* (2010), *Il desiderio di essere come tutti* (2013), *Momenti di trascurabile infelicità* (2015); Tommaso Pincio, *Lo spazio sfinito* (2000), *Hotel a zero stelle. Inferni e paradisi di uno scrittore senza fissa dimora* (2011), *Pulp Roma* (2012); Michele Prisco, *Una spirale di nebbia* (1966), *La provincia addormentata* (1969); Christian Raimo, *Latte* (2001), *Il peso della grazia* (2012); Fabrizia Ramondino, *In viaggio* (1995), *L'isola riflessa* (1998); Ermanno Rea, *Mistero napoletano. Vita e passione di una comunista negli anni della guerra fredda* (1995), *La dismissione* (2002), *La comunista. Due storie napoletane* (2012); Tiziano Scarpa, *Occhi sulla graticola* (1996), *Stabat Mater* (2008), *Le cose fondamentali* (2014), *Il brevetto del geco* (2016); Clara Sereni, *Casalinghitudine* (1987), *Manicomio primavera* (1989), *Eppure* (1995), *Il lupo mercante* (2007), *Una storia chiusa* (2012), *Via Ripetta 155* (2015); Domenico Starnone, *Ex cattedra* (1987), *Il salto con le aste* (1989), *Fuori registro* (1991), *Eccesso di zelo* (1993), *Denti* (1994), *Via Gemito* (2000), *Prima esecuzione* (2007), *Autobiografia erotica di Aristide Gambìa* (2011), *Lacci* (2014), *Scherzetto* (2016); Susanna Tamaro, *La testa tra le nuvole* (1989), *Per voce sola* (1991), *Va' dove ti porta il cuore* (1994), *Ascolta la mia voce* (2006), *Ogni angelo è tremendo* (2013); Chiara Valerio, *Fermati un minuto a salutare* (2007), *Almanacco del giorno prima* (2014), *Storia umana della matematica* (2016); Giorgio Vasta, *Il tempo materiale* (2008), *Spaesamento* (2010); Sandro Veronesi, *Venite venite B-52* (1995), *Caos calmo* (2006), *Brucia Troia* (2007), *Terre rare* (2014); Simona Vinci, *Dei bambini non si sa niente* (1997), *Brother and Sister* (2003).

References

Alfonzetti, G. (2018). Il dialetto 'molesto' in Elena Ferrante. In Marcato, G. (ed.), *Dialetto e società*. Padova: Cleup, 303-314.

Andreoli, R. (1887). *Vocabolario napoletano-italiano*. Torino: Paravia.

Benedetti, L. (2012). Il linguaggio dell'amicizia e della città: *L'amica geniale* di Elena Ferrante tra continuità e cambiamento, *Quaderni d'italianistica*, 33(2), 171-187.

Bolasco, S. (2010). *TaLTaC2.10 Sviluppi, esperienze ed elementi essenziali di analisi automatica dei testi*. Milano: LED.

Baudino, M. (2005). Il «reo» non confesso, *La Stampa*, Torino, 17 January 2005, 27.

Caldwell, L. (2012). Imagining Naples: The Senses of the City. In Bridge, G. and Watson, S. (eds.), *The New Blackwell Companion to the City*. Malden: Wiley-Blackwell, 337-346.

Cavanaugh, J. R. (2016). Indexicalities of Language in Ferrante's Neapolitan Novels: Dialect and Italian as Markers of Social Value and Difference. In Russo Bullaro, G. and Love, S. (eds.), *The Works of Elena Ferrante. Reconfiguring the margins*. New York: Palgrave Macmillan, 45-70.

Ceccoli, V. C. (2017). On Being Bad *and* Good: My Brilliant Friend Muriel Dimen, *Studies in Gender and Sexuality*, 18(2), 110-114.

Chemotti, S. (2009). *L'inchiostro bianco. Madri e figlie nella narrativa italiana contemporanea*. Padova: Il Poligrafo.

Cortelazzo, M.A. (2017). Studiosi dall'Europa per tracciare il profilo di Elena Ferrante, *Il Mattino di Padova*, Padova, 8 September 2017, 43.

Cortelazzo, M.A. and Tuzzi, A. (2017). Sulle tracce di Elena Ferrante: questioni di metodo e primi risultati. In Palumbo, G. (ed.), *Testi, corpora, confronti interlinguistici: approcci qualitativi e quantitativi*. Trieste: EUT Edizioni Università di Trieste, 11-24.

Cortelazzo, M.A., Mikros, G.K. and Tuzzi A. (2018). *Profiling Elena Ferrante: a Look Beyond Novels*. In Iezzi, D.F., Celardo L., Misuraca M. (eds.), *Jadt '18. Proceedings of the 14th International Conference on Statistical Analysis of Textual Data*. Roma: UniversItalia, 165-173.

Cortelazzo, M., Nadalutti, P. and Tuzzi, A. (2013). Improving Labbé's intertextual distance: testing a revised version on a large corpus of Italian literature, *Journal of Quantitative Linguistics*, 20(2), 125-152.

Cortelazzo, M. A., Nadalutti, P., Ondelli, S. and Tuzzi, A. (2018). Authorship Attribution and Text Clustering in Contemporary Italian Novels: Does Elena Ferrante's and Domenico Starnone's regional origin play a role? In: Wang, L., Köhler, R., Tuzzi, A. (eds.), *Scructures, properties, and interrelations. Selected papers from Qualico 2016*. Lüdenscheidt: RAM Verlag, 1-14.

De Santis, R. (2017). Le prove sono nella letteratura. "Elena Ferrante è Starnone", *La Repubblica*, Roma, 8 September 2017, 43.

Dow, G. (2016). The 'biographical impulse' and pan-European women's writing. In Batchelor J. and Dow G. (eds.), *Women's Writing, 1660-1830: Feminisms and Futures*. London: Palgrave Macmillan, 193-213.

Falotico, C. (2015). Elena Ferrante. Il ciclo dell'Amica geniale tra autobiografia, storia e metaletteratura, *Forum Italicum*, 49(1), 92-118.

Ferrante, E. (1992). *L'amore molesto*. Roma: E/O.

Ferrante, E. (2002). *I giorni dell'abbandono*. Roma: E/O.

Ferrante, E. (2006). *La figlia oscura*. Roma: E/O.

Ferrante, E. (2007). *La spiaggia di notte*. Roma: E/O.

Ferrante, E. (2011). *L'amica geniale. Infanzia, adolescenza*. Roma: E/O.

Ferrante, E. (2012). *Storia del nuovo cognome. L'amica geniale volume secondo*. Roma: E/O.

Ferrante, E. (2013). *Storia di chi fugge e di chi resta. L'amica geniale volume terzo*. Roma: E/O.

Ferrante, E. (2014). *Storia della bambina perduta. L'amica geniale volume quarto*. Roma: E/O.

Ferrante, E. (2016). *La Frantumaglia*. Roma: E/O.

Galella, L. (2005). Ferrante-Starnone. Un amore molesto in via Gemito, *La Stampa*, Torino, 16 January 2005, 27.

Galella, L. (2006). Ferrante è Starnone. Parola di computer, *L'Unità*, Roma, 23 November 2006.

Gatti, C. (2016). Elena Ferrante, le «tracce» dell'autrice identificata, *Il Sole 24 Ore – Domenica*, Milano, 2 October 2016, 1-2.

Gatto, S. (2006). Starnone-Ferrante: quando il senso di colpa genera doppi, *Lo Specchio di carta. Osservatorio sul romanzo italiano contemporaneo*, 28 October 2006 (www.lospecchiodicarta.it).

Gatto, S. (2016). Una biografia, due autofiction. Ferrante-Starnone: cancellare le tracce, *Lo Specchio di carta. Osservatorio sul romanzo italiano contemporaneo*, 22 October 2016 (www.lospecchiodicarta.it).

Labbé, C. and Labbé, D. (2001). Inter-textual distance and authorship attribution. Corneille and Molière, *Journal of Quantitative Linguistics*, 8(3), 213-231.

Lee, A. (2016). Feminine Identity and Female Friendships in the 'Neapolitan' Novels of Elena Ferrante, *British Journal of Psychotherapy*, 32(4), 491-501.

Librandi, R. (*in press*). Una lingua silenziosa: immaginare il dialetto negli scritti di Elena Ferrante. In Jamrozik E. and Tylusinska-Kowalska, A. (eds.), *Dal monologo al polilogo: l'Italia nel mondo. Lingue, letterature e culture in contatto, Atti del Convegno (Varsavia 6-8 aprile 2017).*

OrphAnalytics (2016). Determination by stylometry of the probable author of

the Ferrante corpus: Domenico Starnone, 11 October 2016, http://www.orphanalytics.com

Ricciotti, A. (2016). Un confronto tra Elena Ferrante e Anna Maria Ortese: la città di Napoli, la fuga, l'identità, *Zibaldone. Estudios italianos de La Torre del Virrey*, 4(2), 111-122.

Russo Bullaro, G. and Love, S. V. (2016, eds). *The Works of Elena Ferrante. Reconfiguring the Margins*. New York, Palgrave Macmillan.

Santagata, M. (2016). Elena Ferrante è…, *La lettura – Corriere della Sera*, Milano, 13 March 2016, 2 and 5.

Starnone, D. (2000). *Via Gemito*. Milano: Feltrinelli.

Tuzzi, A. (2010). What to put in the bag? Comparing and contrasting procedures for text clustering. *Italian Journal of Applied Statistics/Statistica Applicata*, 22(1), 77-94.

Tuzzi, A. and Cortelazzo, M.A. (2018). What is Elena Ferrante? A comparative analysis of a secretive bestselling Italian writer, *Digital Scholarship in the Humanities* (online first 19 January 2018 fqx066, https://doi.org/10.1093/llc/fqx066).

Zingarelli (2015). *Lo Zingarelli 2016. Vocabolario della lingua italiana*. Bologna: Zanichelli.

Elena Ferrante: A Virtual Author

Maciej Eder
Pedagogical University of Kraków and Polish Academy of Sciences, Poland

Abstract

The present study scrutinizes the novels by Elena Ferrante, in order to discover the actual writer hidden behind the pseudonym. Rather than simply reopen the authorship question, however, the paper attempts at testing the stability of the authorial signal in the works by "Ferrante", whoever the actual author might be. To address the research question, a network of 150 novels and their stylistic similarities has been computed using the Bootstrap Consensus Network method. A list of authors most similar to "Ferrante", including Domenico Starnone at the first place, was then analyzed using the technique Rolling Classify, which was designed to detect local stylistic idiosyncrasies in literary texts. The series of Rolling Classify tests – performed independently for the novels by both Ferrante and Starnone – allows for formulating general observations. The overall picture confirms, with a few exceptions, that Starnone and Ferrante can be told apart, which, in turn, seems to be a strong argument in favor of the virtual author hypothesis. Apparently, Domenico Starnone demonstrates, particularly in his late works, the ability to differentiate his own stylistic profile and the voice of his *alter ego*.

Introduction

Elena Ferrante, an Italian writer who's novels have gained international recognition, has been publishing her (his? their?) work for almost three decades know. Seven novels, and recently a weekly column in *The Guardian* – the considerable amount of word-class literature remains, in a way, anonymous. It is true that we know the pseudonym "Elena Ferrante", but the actual name of the author remains unknown to the public. Or, rather, one should say that

the author's name remained unknown until recently. In a few journal articles, Domenico Starnone has been suggested as the actual author behind the famous pseudonym (Galella, 2005; Gatto, 2016). In recently conducted studies involving state-of-the-art statistical methodology, these findings have been further scrutinized (Cortelazzo and Tuzzi, 2017; Cortelazzo, Mikros and Tuzzi, 2018; Tuzzi and Cortelazzo, 2018). Starnone, born in 1943, is a prolific Italian writer and journalist, who has published a dozen of well-received novels and a number of minor works. Interestingly, another suggested author of the novels by "Ferrante" is Anita Raja – a translator of German literature and Domenico Starnone's wife (Gatti, 2016).

The present study is not intended to reopen the above authorship question, although the Starnone hypothesis will play an important role here. Instead, the paper attempts at testing the stability of the authorial signal in the works by "Ferrante", whoever the actual author turns out to be. Rather than simply unmasking the name, the paper will test whether – and if yes, then to which extent – the unmasked author's own novels differ stylistically from the works published as "Ferrante". The research question, then, is as follows: do we deal with two distinct yet coherent "authorial" profiles of a writer and his/her second persona? The above question will be assessed using authorship attribution techniques, and operationalized in the form of the following hypothesis: if a machine-learning classifier is able to tell apart two non-identical yet closely related authorial fingerprints – one belonging to "Ferrante", and the other to an actual Italian writer – the existence of two stylistic personae will be claimed probable.

Shortlisting the candidates

An authorship attribution investigation should start with selecting the texts by known authors who could have written the anonymous text in question. In the case of Ferrante, it is more than risky to assume that the list of possible "candidates" can be exhausted, even if Domenico Starnone is somewhat more likely as a "candidate" than any other Italian writer. A natural strategy to attack such an open-set attribution case, is to collect a reasonably high number of works by contemporary authors, in order to perform a large-scale screening via a series of stylometric tests, followed by identifying possible "candidates". In the present study, a corpus of 150 Italian 20th-century novels by 40 authors – out of which 7 are penned by Ferrante – has been used. The corpus has been compiled by Michele Cortelazzo and Arjuna Tuzzi and used in previous studies on Ferrante (Cortelazzo and Tuzzi, 2017; Cortelazzo *et al.*, 2018; Tuzzi and Cortelazzo, 2018).

Since the goal is to narrow the list of potential "candidates" rather than to solve the attribution case – this is a classification scenario in which recall is more important than precision – a relatively simple exploratory method will be applied. It is true that sophisticated machine-learning techniques, such as Support Vectors Machines, prove extremely efficient to solve multidimensional problems (James *et al.*, 2013). However, in authorship attribution relatively simple distance-based methods seem to perform sufficiently well (Jockers and Witten, 2010), especially when the number of authorial classes is high (Luyckx and Daelemans, 2011).

To shortlist the possible authors of Ferrante's novels, the Bootstrap Consensus Network (BCN) method has been used (Eder, 2017). It is an enhanced variant of cluster analysis, designed to assess a given corpus several times using different sets of features (here, different vectors of most frequent words, ranging from 100 to 1,000). In each iteration, a stylometric distance-based test for textual similarity is applied. The goal is to identify the texts stylistically related one to another.

Essentially, each text in a corpus can be characterized as being more or less similar (stylometrically) to all the other texts. This implies that for each text, there exists its stylistic *nearest neighbor*. A geometric interpretation of the above intuition is a multidimensional space (the number of dimensions being as high as the number of features to be measured) in which texts are represented as points; the closer the geometric distance between two analyzed points, the more similar they are. Hence, the proximity between texts can be defined geometrically, via the notion of a *distance measure*. The measure used in the present study is Manhattan distance applied to scaled (z-scored) word frequencies, which is then divided by the number of dimensions (i.e. analyzed words). This measure is also referred to as Burrows's Delta distance (Burrows, 2002).

Each iteration of the BCN method results in a list of nearest neighbor relations between texts for a given number of features – the first "snapshot" measures the relations for 100 most frequent words, the second iteration captures 200 words, then 300, 400 and so forth, all the way to 1,000 words. The next step involves combining particular "snapshots", and mapping them onto a network. The connections of the network are nearest neighbor relations between texts. More precisely, every single text is assigned three connections – the nearest neighbor of a given text and its two runner-ups. Consequently, the BCN procedure shares some similarities with the *k*-NN classifier, which makes a classification decision using *k* nearest neighbors of a given text. The non-parametric nature of *k*-NN (James *et al.*, 2013, p. 104) is, by definition, also a feature of BCN.

A network produced by the above technique is sufficiently informative *per se*; however, since the list of network connections might be somewhat difficult

to inspect by a naked eye, the network can be additionally visualized using one of the force-directed layouts. In such a case, the interpretation step involves manual inspection of the nodes (i.e. texts) that usually lump into clusters, which might be further lumped into larger groups of nodes. Human interpretation of the emerging clusters makes the method rather straightforward to use; this is a common feature of several explanatory methods.

To perform all the computational analyses reported in this paper, including the stylometric networks, the stylometric R package "Stylo" has been used (Eder, Rybicki and Kestemont, 2016). The final networks have been visualized using the software Gephi and its built-in "Force2 Atlas" algorithm to produce the network's layout (Bastian, Heymann and Jacomy, 2009).

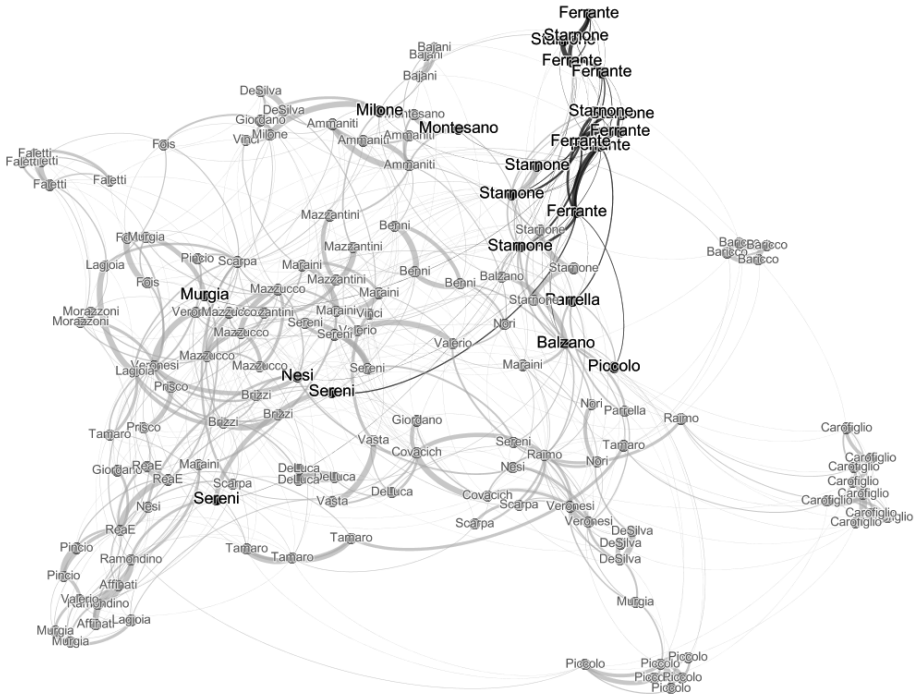

Figure 1. Bootstrap consensus network of 150 contemporary Italian novels. Nearest neighbor similarities between texts are represented as network connections. The texts similar to the novels by Ferrante are marked in black color.

A consensus network of the corpus of 150 contemporary Italian novels is shown in Fig. 1. Apart from the fact that the novels formed a "map" of the Italian literature that can be further interpreted in terms of the emerging clusters, the most relevant to this study are, at the first place, the neighbors of Ferrante's novels. In particular, one can see the proximity of Ferrante and Starnone. It should be emphasized that Starnone is by far the most similar author, out of all the writers represented in Fig. 1. Essentially, each of the seven books by Ferrante is robustly connected with one of the novels by Starnone. There are also a few other authors, however, who appear to have occasionally established network connections with Ferrante – connections that should be interpreted in terms of stylistic similarities. These novelists are as follows: Marco Balzano, Rosella Milone, Giuseppe Montesano, Valeria Parrella, Francesco Piccolo, Clara Sereni, and, to a lesser extent, Michela Murgia and Edoardo Nesi. Certainly, the above list contains a rather heterogenous constellation of female and male Italian writers, nevertheless all of them will be considered as potential candidate authors than might be hidden under the pseudonym Elena Ferrante.

Ferrante in a moving window

The experiment reported in the previous section shows that Domenico Starnone might be indeed Ferrante's *alter ego*, let alone a few other yet less evident candidate authors. To further explore possible stylistic relations between the shortlisted novels, a substantially different experimental setup will be applied, namely Rolling Stylometry, a relatively new method to examine (anonymous) texts that are split into smaller parts and then assessed sequentially (Eder, 2015). It relies on a general intuition that stylometric signals do not need to be equally strong in different parts of a given text. Suffice it to say that narrative parts of a novel will form a different stylometric profile than the passages containing dialogues. Similarly, larger amounts of intertextuality, e.g. extensive quotations, should be noticeable through the lenses of sequential methods – when each segment of a novel is tested as a separate (yet sequentially ordered) entity.

The reason of applying Rolling Classify to the novels by Ferrante is, firstly, a suspicion that "Ferrante" might be a composite of more than one actual author, and secondly, the assumption that a bird's eye view at the large network of literature should be supplemented by a magnifying glass perspective, in which every single chunk of a text matters.

The method in question is a supervised machine-learning classifier, which means that it requires a *training set*, or a collection of manually selected texts (of known authorship) representative for their authorial classes, and a *test set*

that contains the anonymous text to be examined. The difference between the usual supervised setup and the Rolling Classify method is that the latter splits the input text into equal-sized blocks, and then populates the *test set* with the discrete text chunks, while keeping their original order. The method is supported with compact visualization that shows the original text – or, to be precise, the sequence of its chunks – in the form of a color stripe. The bottom part of the stripe reflects the decision of the classifier: the thicker the stripe, the more robust the classification.

In the present study, all the sequential experiments were performed using the same set of hyperparameters. The training set contained 30 novels by seven writers shortlisted in the previous experiment: Marco Balzano, Rossella Milone, Giuseppe Montesano, Valeria Parrella, Francesco Piccolo, Clara Sereni, and Domenico Starnone. In a set of independent tests, each novel by Ferrante was split using a sliding window of the length of 5,000 words; the window was moving forward through the original text at the pace of 500 words. Frequencies of 100 most frequent words were used as stylometric features, Support Vector Machine algorithm as a classifier.

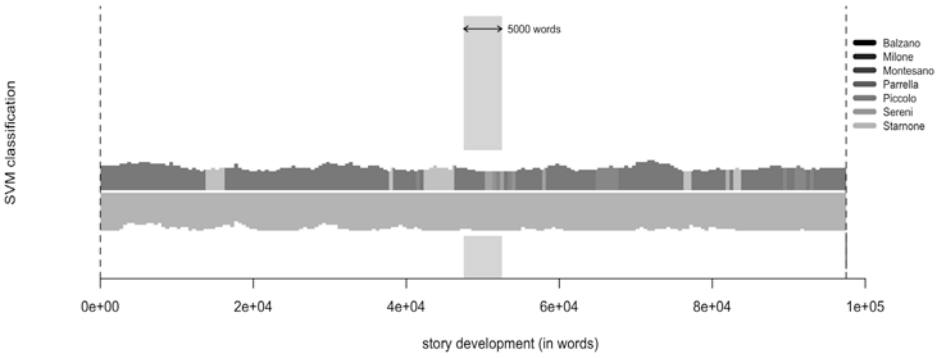

Figure 2. *L'amica geniale* by Ferrante, contrasted sequentially with novels by 7 shortlisted authors. The Rolling Classify technique, with SVM as a classifier, applied to 100 most frequent words as features.

The results for *L'amica geniale* by Ferrante (2011) are shown in Fig. 2. The uniform stripe indicates no stylistic takeovers – the entire novel in all its segments is classified as written by Domenico Starnone. The picture for all the remaining novels by Ferrante turned out to be simply identical: with no single exception all the segments in all of these novels were ascribed to Starnone. The evidence is strong here.

A virtual author?

Even if authorship problems can never be simply claimed *solved* – due to the very nature of statistical inference – the hypothesis of the identity of Starnone and Ferrante seems to have solid empirical evidence. Interesting as it is, however, the procedure of unmasking the actual author's name does not tell much about the novels themselves, nor does it contribute to our understanding of the writing process. Even if it is usually good to know who the author of a literary work is, one would be more interested in discovering to which extent a writer can intentionally change his/her style. Assuming that Ferrante is Starnone, one would like to know what is the relation between the two personae – when Starnone decides to write under the pseudonym, is he any different from Starnone writing as himself?

Certainly, the problem of a double identity is not new. One of the best known examples in the European literature is Romain Gary (1914–1980), a French novelist who at a certain stage of his career decided to change his persona, and started publishing under the pseudonym Émile Ajar. He is the only author to have won twice the Prix Goncourt, which in fact can be awarded only once to an author. It could have happened because his second persona had not been unmasked. Stylometric evidence based on rhythmical patterns of his prosody suggests that his two authorial profiles are indeed different, even if the signal is not entirely clear (Pawłowski, 1996).

A change of one's own authorial profile does not need to be intentional. Worth mentioning is a natural stylistic drift over time (Stamou, 2008), clearly noticeable in the writings of Henry James (Hoover, 2007). Interestingly, one of the first applications of stylometric methodology was a study aimed at tracing the development of Plato's authorial style during his lifespan (Lutosławski, 1897). A change in one's style might be due to some health issues, e.g. dementia, as in the case of Agatha Christie (Le *et al.*, 2011), but it can also happen when a writer hires a secretary to partially take over, as has been observed in the writings of Francis Bacon (Reynolds, Schaalje and Hilton, 2012).

Relevant to the present study is a variant of the above phenomenon reported by Kestemont and his colleagues, who analyzed two medieval Latin visions traditionally ascribed to Hildegard of Bingen (Kestemont, Moens and Deploige, 2013). The scholars have observed that works written collaboratively by Hildegard and her secretary Guibert of Gembloux form a discrete profile, which is not a simple combination of both Hildegard's and Guibert's styles. Rather, we deal here with a virtual "third author". A theoretical framework for the existence of virtual authors has been provided by the Synergy Hypothesis (Pennebaker, 2011), according to which the style of a collaboratively written work can resemble a predominant author's style, but it can also be unlike either of one of the styles that the collaborating authors would produce on their own.

The next sections will try to examine whether the relation Starnone vs. Ferrante anyhow suggests the existence of a virtual author (Ferrante) that would be significantly different from Starnone himself. This question is in fact multifaceted, because it also involves other issues, e.g. the question of gender: does Ferrante exhibit more features typical to female writers than Starnone? Was the actual author successful in dissolving his own dialectal idiosyncrasies? Only a fraction of these and similar questions can be undertaken in the present study.

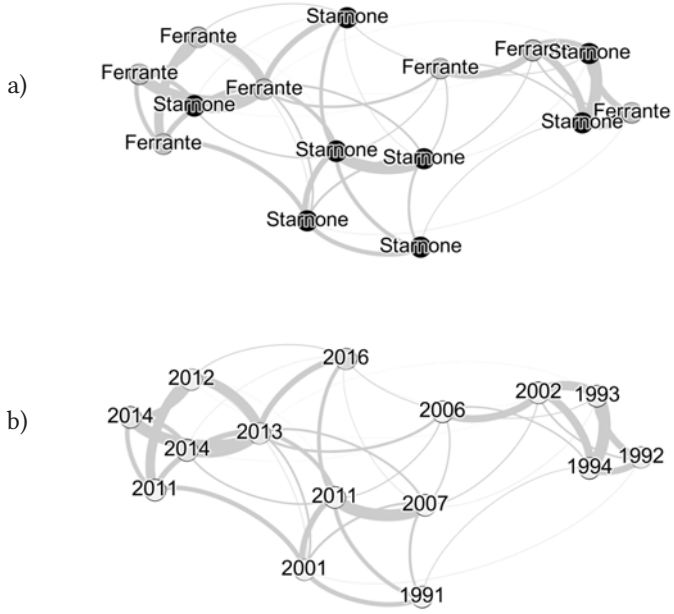

Figure 3. Bootstrap Consensus Network of the similarities between the novels penned by Ferrante and Starnone: (a) the distinction between the two authors: no clear pattern emerges, (b) the existence of a temporal signal, which seems to be stronger than two authorial voices.

The analysis will start with a small-scale consensus network focusing exclusively on the relations between the novels by Starnone and Ferrante. As can be seen in Fig. 3a, the two authorial classes tend to form some clusters, but the distinction is far from being clear-cut. Rather, the novels seem to follow a *temporal* pattern, as evidenced in Fig. 3b. Even if preliminary, the results revealed one additional – and relatively strong – signal to be aware of: the assumed distinct voices of Starnone and his *alter ego* might be overshadowed by the author's general stylistic drift over time.

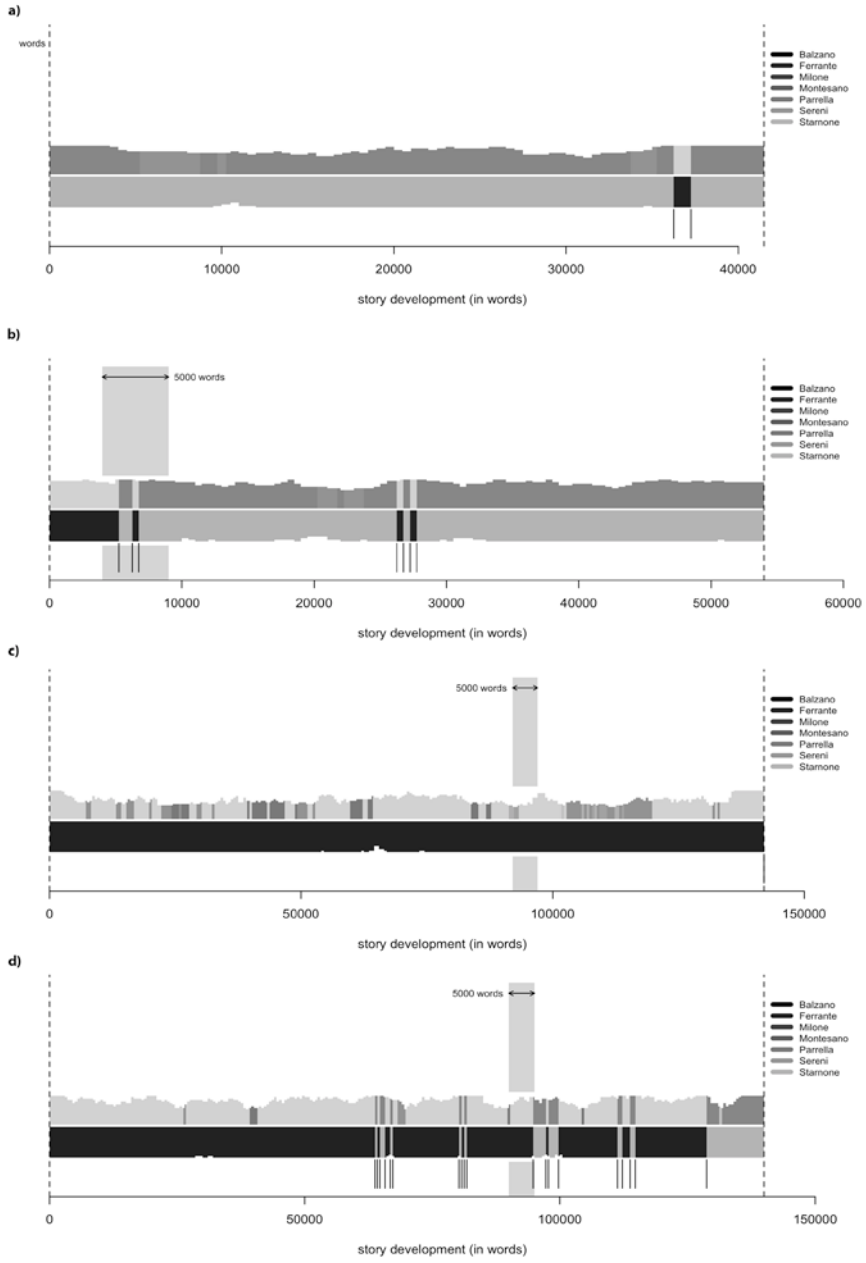

Figure 4. Novels by Elena Ferrante assessed sequentially via the Rolling Classify technique: (a) *L'amore molesto* (1992), (b) *I Giorni dell'abbandono* (2002), (c) *Storia del nuovo cognome* (2012), (d) *Storia della bambina perduta* (2014). The training set contained 30 novels by 7 authors, including both Starnone and Ferrante.

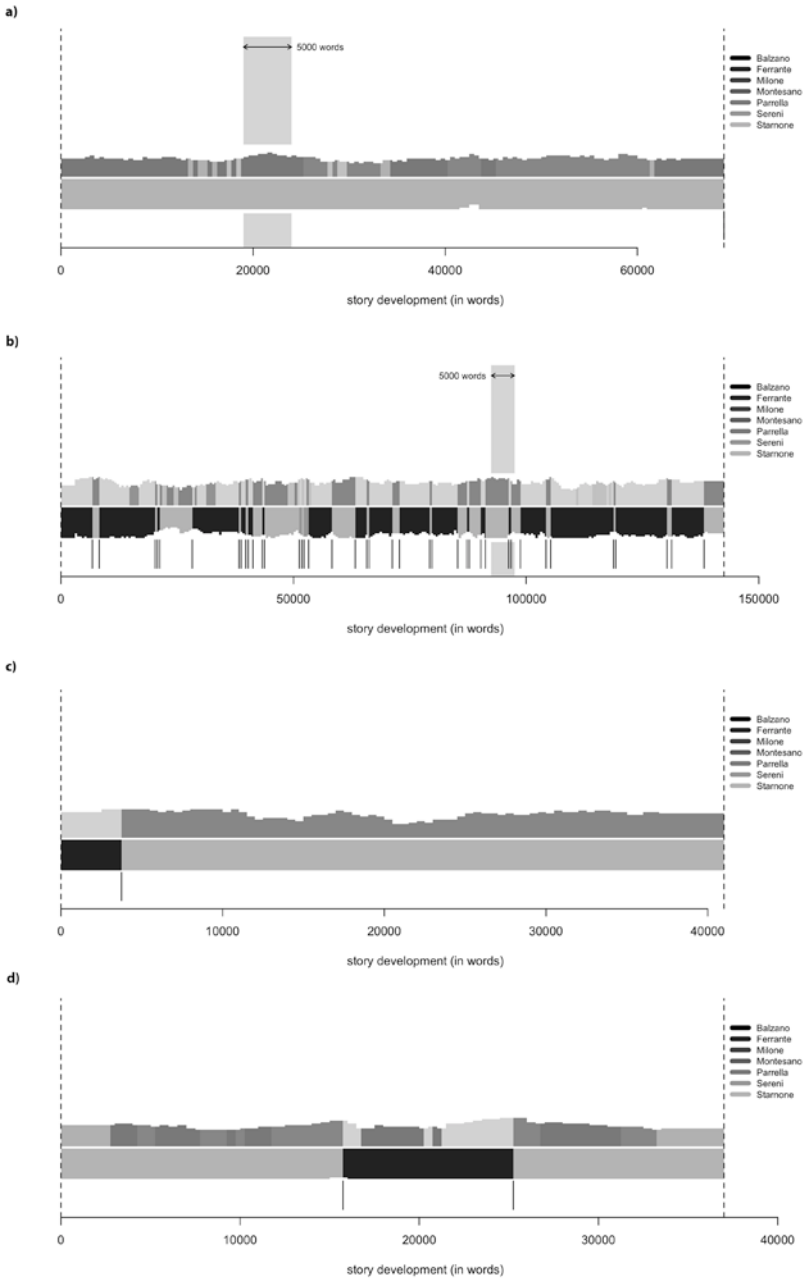

Figure 5. Novels by Domenico Starnone assessed sequentially via the Rolling Classify technique: (a) *Il salto con le aste* (1989), (b) *Via Gemito* (2000), (c) *Prima esecuzione* (2007), *Lacci* (2014). The training set contained 36 novels by 8 authors, including both Starnone and Ferrante.

To further explore the above issue, an additional series of tests have been performed. Again, the Rolling Stylometry technique has been used; this time, however, the training set contained the works of the seven authors as mentioned above *and* the works by Ferrante. Needless to say, the novels by Starnone and Ferrane were marked as two distinct classes. The research question is as follows: will the classifier recognize the virtual Ferrante? The hyperparameters were the same as in the previous setup: Support Vectors Machine, 100 most frequent words, the window of 5,000 words, sliding at a pace of 500 words.

The results of the experiments are shown in Fig. 4a–d (four novels out of seven); the analyzed novels by Ferrante are ordered chronologically. Arguably, a clear pattern appears: while the early novels show little similarity with the assumed virtual "Ferrante", the late works are assigned to this class with more and more confidence of the classifier. Almost all of the segments of *L'amore molesto* from 1992 (Fig. 4a) are classified as "Starnone", with an exception of a relatively short passage at the end of the novel. The voice of the virtual "Ferrante" is more noticeable in *I Giorni dell'abbandono* from 2002 (Fig. 4b), this time at the beginning of the novel. In *La figlia oscura* (2006) the share of segments by "Ferrante" is roughly equal to those of "Starnone". In the novel *L'amica geniale. Infanzia, adolescenza* (2011) the style of "Ferrante" becomes predominant, which is even more visible in *Storia del nuovo cognome* published 2012 (Fig. 4c). This novel is a triumph of the virtual author: all of the segments have been attributed to the class "Ferrante". Even if in two later novels – *Storia di chi fugge e di chi resta* (2013) and *Storia della bambina perduta* (2014) – the signal is somewhat blurry (Fig. 4d), the predominant voice of the virtual author cannot be denied.

No matter how striking the obtained results are, they might as well contain some bias. It is true that the evidence is strong, and supports the hypothesis that the "authorial" profile of Ferrante has been gradually emerging, to become predominant in the late novels. However, before such a claim could be considered valid, one would have to test if the original authorial voice of Starnone remained equally stable during the same period of time. Indeed, an alternative hypothesis, claiming that Starnone simply turned into Ferrante (no matter under which name he was publishing), cannot be rulled out. Moreover, as evidenced in the previous test (Fig. 3b), a general temporal signal in the dataset has been already observed.

To validate the results obtained for the Ferrante's novels, then, another set of tests needs to be performed, this time focused on the novels penned by Starnone. Certainly, the same set of parameters as in the previous experiment needs to be kept. The results are plotted in Fig. 5a–d (four novels out of ten). They indeed suggest the existence of Starnone's own authorial signal, non-identical with the voice of the virtual Ferrante, even if the picture is not as clear-cut as one might have expected. In particular, when the novels are ordered chrono-

logically, a few stylistic peculiarities are noticeable. To start with the earliest works: the segmented text of *Ex cattedra* (1987) reveals a pure voice of Starnone, as does *Il salto con le aste* published in 1989 (Fig. 5a). *Fuori registro* (1991) is roughly similar to the first two novels, except that it contains a passage – a few thousand words at the beginning – classified as "Ferrante". Next comes *Eccesso di zelo* (1993), in which, again, the profile of Starnone prevails over two minor passages by "Ferrante". The situation changes with *Denti* (1994), having roughly the same amount of text classified to either class, and *Via Gemito* from 2000 (Fig. 5b), where the classifier hardly recognizes any "Starnone". Not only is this novel assigned to "Ferrante", but it also contains a few (minor) chunks that are attributed to Clara Sereni (!), which apparently undermines the hypothesis of Starnone's ability to differentiate the voices. On the other hand, the next novel, namely *Prima esecuzione* (2007), neutralizes the above concern (Fig. 5c), since the voice of Starnone is clear again, with a relatively small amount of "Ferrante" at the beginning of the book. The author continues to have his own stylistic profile in *Autobiografia erotica di Aristide Gambìa* (2011), even if a few random chunks attributed to either Elena Ferrante or Clara Sereni might occasionally appear.

Switching between two different personae must be non-trivial even for a world-class writer, as evidenced in the next novel, *Lacci* (2014), contaminated by a lengthy passage by "Ferrante" (Fig. 5d). However, the recently published work *Scherzetto* (2016) appears to be, again, the author's successful attempt to mute "Ferrante" in himself: apart from a few marginal text chunks, the vast majority of passages are robustly attributed to Starnone.

Even if some of the above outcomes do not allow for making any definite interpretations, a significant majority of the novels by both Ferrante and Starnone tend to keep a relatively clear voice of their respective "authors". It turns out that the literary mystification of an Italian writer from Naples was, generally, more than successful, at least when approached via stylometric methodology.

Conclusions

The present study was aimed at scrutinizing the authorship of the novels published under the pseudonym Elena Ferrante. As could be demonstrated, the already suggested hypothesis of Domenico Starnone's authorship of the novels in question was difficult to falsify. The main goal, however, was to go beyond the simple question of authorship, and to test how successful (stylistically) was the actual author hidden under the pseudonym, in comparison to his/her own writings.

The series of Rolling Classify tests – performed independently for the novels by both Ferrante and Starnone – allows for formulating general observations. The overall picture confirms, with a few exceptions, that Starnone and Ferrante can be told apart, which, in turn, seems to be a strong argument in favor of the virtual author hypothesis. Apparently, Domenico Starnone demonstrates, particularly in his late works, the ability to differentiate his own stylistic profile and the voice of his *alter ego*.

The phenomenon that definitely deserves further investigation is the passages of the works by Starnone which have been wrongly classified as being written by Clara Sereni (in *Via Gemito*) or Marco Balzano (in *Lacci*). On the one hand, these misattributions can be explained as, simply, wrong decisions of the insufficiently trained classifier – such instances are referred to as "false positives" in machine learning. On the other hand, however, we might deal here with local stylistic idiosyncrasies in the novels themselves. Such instances of the authorial signal locally overshadowed by apparently someone else's fingerprint have been already observed in the novel *To Kill a Mockingbird* by Harper Lee (Eder and Rybicki, 2016). The recognized traces of Truman Capote could hardly be due to the mixed authorship of the novel. More likely, we deal here with the phenomenon of intertextual voices that might locally shine through the (usually opaque) original author's fingerprint. Arguably, a similar phenomenon might have occurred in *Via Gemito* and in *Lacci* by Domenico Starnone.

The obtained results are striking because, firstly, they show a gradual – and successful – development of a virtual author, barely visible in the early works by "Ferrante", and predominant in the late ones. Secondly, the results are a meaningful contribution to the above-mentioned Synergy Hypothesis (Pennebaker, 2011). Not only can it take the form of a stylistic combination of two different authorial voices, but also: two different stylistic personae hidden behing one actual author.

References

Bastian, M., Heymann, S. and Jacomy, M. (2009). Gephi: An open source software for exploring and manipulating networks. In *Proceedings of the Third International ICWSM Conference*. San Jose, 361-362.

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

Cortelazzo, M.A., Tuzzi, A. (2017). Sulle tracce di Elena Ferrante: Questioni di metodo e primi risultati. In Palumbo G. (ed.), *Testi, corpora, confronti interlinguistici: Approcci qualitativi e quantitative*. Trieste: Edizioni Università di Trieste, 11-25.

Cortelazzo, M.A., Mikros, G.K. and Tuzzi A., *Profiling Elena Ferrante: a Look Beyond Novels*. In Iezzi, D.F., Celardo L., Misuraca M. (eds.), *Jadt '18. Proceedings of the 14th International Conference on Statistical Analysis of Textual Data*. Roma: UniversItalia, 165-173.

Cortelazzo, M.A., Nadalutti, P., Ondelli, S. and Tuzzi, A. (2018). Authorship Attribution and Text Clustering in Contemporary Italian Novels: Does Elena Ferrante's and Domenico Starnone's regional origin play a role? In Wang, L., Köhler, R. and Tuzzi, A. (eds.), *Scructures, properties, and interrelations. Selected papers from Qualico 2016*. Lüdenscheidt: RAM Verlag, 1-14.

Eder, M. (2015). Rolling stylometry, *Digital Scholarship in the Humanities*, 31(3), 457-469.

Eder, M. (2017). Visualization in stylometry: Cluster analysis using networks, *Digital Scholarship in the Humanities*, 32(1), 50-64.

Eder, M. and Rybicki, J. (2016). Go set a watchman while we kill the mockingbird in cold blood, with cats and other people. In *Digital Humanities 2016: Conference Abstracts*, 184-186. Kraków: Jagiellonian University & Pedagogical University (http://dh2016.adho.org/abstracts/70).

Eder, M., Rybicki, J. and Kestemont, M. (2016). Stylometry with R: A package for computational text analysis, *R Journal*, 8(1), 107-121 (https://journal.r-project.org/archive/2016/RJ-2016-007/index.html).

Galella, L. (2005). Ferrante-Starnone. Un amore molesto in via Gemito, *La Stampa*, Torino, 16 January 2005, 27.

Gatti, C. (2016). Elena Ferrante: An answer? *The New York Review of Books* (http://www.nybooks.com/daily/2016/10/02/elena-ferrante-an-answer/).

Gatto, S. (2016). Una biografia, due autofiction. Ferrante-Starnone: cancellare le tracce, *Lo Specchio di carta. Osservatorio sul romanzo italiano contemporaneo*, 22 October 2016 (www.lospecchiodicarta.it).

Hoover, D. (2007). Corpus stylistics, stylometry, and the styles of Henry James, *Style*, 41(2), 174-203.

James, G., Witten, D., Hastie, T. and Tibshirani, R. (2013). *An introduction to statistical learning with applications in R*. New York: Springer.

Jockers, M.L. and Witten, D. M. (2010). A comparative study of machine learning methods for authorship attribution, *Literary and Linguistic Computing*, 25(2), 215-223.

Kestemont, M., Moens, S. and Deploige, J. (2013). Collaborative authorship in the twelfth century: A stylometric study of Hildegard of Bingen and Guibert of Gembloux, *Literary and Linguistic Computing*, *28*, 1-15 (https://doi.org/doi:10.1093/llc/fqt063).

Le, X., Lancashire, I., Hirst, G. and Jokel, R. (2011). Longitudinal detection of dementia through lexical and syntactic changes in writing: A case study of three British novelists, *Literary and Linguistic Computing*, 26(4), 435-461.

Lutosławski, W. (1897). *The origin and growth of Plato's logic: With an account of Plato's style and of the chronology of his writings*. London: Longmans, Green & Co.

Luyckx, K. and Daelemans, W. (2011). The effect of author set size and data size in authorship attribution, *Literary and Linguistic Computing*, 26(1), 35-55.

Pawłowski, A. (1996). *Séries temporelles en linguistique: Avec application à l'attribution de textes, Romain Gary et Emile Ajar*. Lausanne: Slatkine.

Pennebaker, J.W. (2011). *The secret life of pronouns: What our words say about us*. New York: Bloomsbury Press.

Reynolds, N.B., Schaalje, G.B. and Hilton, J.L. (2012). Who wrote Bacon? Assessing the respective roles of Francis Bacon and his secretaries in the production of his English works, *Literary and Linguistic Computing*, 27(4), 409-425 (https://doi.org/10.1093/llc/fqs020).

Stamou, C. (2008). Stylochronometry: Stylistic development, sequence of composition, and relative dating, *Literary and Linguistic Computing*, 23(2), 181-199.

Tuzzi, A. and Cortelazzo, M.A. (2018). What is Elena Ferrante? A comparative analysis of a secretive bestselling Italian writer, *Digital Scholarship in the Humanities* (online first 19 January 2018 fqx066, https://doi.org/10.1093/llc/fqx066).

Thesaurus-Based Semantic Similarity Judgments: A New Approach to Authorial Similarity?

Patrick Juola
Duquesne University, Pittsburgh, PA USA juola@mathcs.duq.edu

Abstract

Authorship attribution is most commonly done by feature comparison as a form of classification or cluster analysis. The typical feature sets used are, by design, low-level features that are difficult to interpret in terms of literary theory, such as letters, words, grammatical categories, or their n-grams. This paper describes the use of semantic categories (as extracted from a conceptual ontology in the form of a published thesaurus) and shows that authors can be categorized on the basis of conceptual and semantic similarity as well.

Introduction

To truly understand the differences between authors, it's not enough simply to say that so-and-so uses "by" twice as often as such-and-such. Most authorship attribution research, however, has focused on classification based on a relatively small number of feature types, including words (primarily frequent words like prepositions and articles), character n-grams, and part of speech tags. This limits the applicability of authorship analysis to develop a richer understanding of the works under study, although the analysis itself may be accurate enough for practical questions.

However, this raises the issue of what other feature sets might be used, both to enhance the accuracy of the analysis, but more importantly to provide high-level guidance to help understand how two authors differ. Literary scholars are not typically interested merely in attributing documents; after all, we

already believe we know who wrote most of the words in the literary canon. Instead, the interest is in interpreting and understanding a distinctive authorial voice. This paper explores the use of concepts (as measured by the vocabulary of a thesaurus) to distinguish between authors. We show that this method is tractable, that it largely agrees with the other scholarship in this volume, but also that it produces meaningful statements about how and why two authors differ.

Background

Questions of authorship and attribution have been around for millennia, often resulting in hundreds of pages of discussion in the scholarly literature when the author and questioned document are sufficiently important. In 1728, for example, 'Captain' Goulding published one of the first known anti-Stratfordian challenges to Shakespeare's authorship of his well-known plays (Friedman and Friedman, 1957, p. 1). The Friedmans alone devote nearly 350 pages to this then 200 year old controversy, but even fifty years later it has not died down. Paul's authorship of the Epistle to the Hebrews was first questioned in 240 AD, but the controversy persists: "It seems to me that much of the evidence regarding authorship of the Pastorals is sufficiently ambiguous that the issue cannot be decided" (Porter, 1995, p. 121).

The possibility of using algorithms and statistics to settle this type of question predates the fields of algorithms and statistics themselves; in the mid-19th century, the great mathematician De Morgan wrote

> I wish you would do this: run your eye over any part of those of St. Paul's Epistles which begin with [paulos] – the Greek, I mean – and without paying any attention to the meaning. Then do the same with the Epistle to the Hebrews, and try to balance in your own mind the question whether the latter does not deal in longer words than the former. […] If St. Paul's Epistles which begin with [paulos] gave 5.428 [letters per word] and the Hebrews gave 5.516, for instance, I should feel quite sure that the Greek of the Hebrews […] was not from the pen of Paul (De Morgan, 1851).

This proposal codifies three aspects of modern authorship attribution. First, that attribution can be done on the basis purely of writing style, "without paying any attention to the meaning". Secondly, authors are assumed to have formalized habits of writing style that persist across different documents. Thirdly, that the presence or absence of these habits can be detected by simple procedures and statistical methods. Mosteller and Wallace (1964) applied these principles to the anonymously published *Federalist Papers* and showed that there were

specific words that were much more common in the writings of one potential author than another. For example, Alexander Hamilton never used the word "whilst" and, by contrast, James Madison never used the Americanized version "while". Perhaps more interestingly, while both authors used the word "by" (after all, a very frequent English word), Hamilton never used the word "by" more frequently than 13 instances per thousand words, while Madison never used it less than 5 per thousand and often as much as 19 per thousand.

Research into different authorship attribution methods has blossomed in the past decade (Juola, 2008; Koppel, Schler and Argamon, 2009; Stamatatos, 2009). Many procedures have been explored, and formal protocols for the analysis of questioned documents have been proposed (Juola, 2015) to help address, among other things, the specific evidentiary needs of the court system.

A good tutorial example is presented in (Binongo, 2003). Binongo examined the authorship of *The Royal Book of Oz*, the 15th book in the *Oz* series, created by L. Frank Baum and continued after his death by Ruth Plumly Thompson. The Royal Book falls into that gap, having been published immediately after Baum's death, but the specific authorship is unclear. Moore (1974, p. 89; cited by Binongo 2003, p. 10) states: "Notes and a fragmentary draft of … The Royal Book of Oz were presumably turned over to a successor, Ruth Plumly Thompson, but no one seems to know exactly how much of this book was really Baum's work…". Binongo collected 14 works of undisputed Baum authorship, another 14 of Thompson's, and the disputed Royal Book. From these books, he extracted the fifty most common words (these are typically short, semantically light words like articles, conjunctions, and prepositions, such as "the", "and", "to", "that", "with", and so forth) and compiled frequency statistics for each word in each work. Using Principal Component Analysis (PCA), he was able to reduce this fifty-dimensional space to its two principal components. When the individual works were plotted, there was a clear visual separation between the works of Thompson (on the left side of the plot) and Baum (on the right). The *Royal Book* also landed on the left side of the plot, in the middle of Thompson's samples. In Binongo's words, "the first [principal component] clearly separates the two authors" [p. 13], and he concludes that "this book is more likely to have been written in Thompson's hand" [p. 14] and that "the statistical analysis in this article reveals that the writing style in the 15th Book of Oz is more compatible with Thompson's than with Baum's".

This analysis is typical of an authorship attribution study:

- Gather writing samples from the author(s) of interest for use as training documents (aka "known documents").
- Extract stylistic features from the training documents.
- Extract the same stylistic features from the test document (aka "unknown document").

- Compare the feature distribution from the test document to the various training documents.
- Authorship of the test document is presumptively attributed to the author whose feature distribution is most similar.

Of course, there are many ways to do this. Similarity, for example, can be assessed via some sort of distance function (Noecker Jr and Juola, 2009) or via a more sophisticated classification technique such as support vector machines (De Vel et al, 2001; Joachims, 2002; Diederich et al., 2003) or decision trees (Quinlan, 1993), or using machine learning methods such as neural networks (Tweedie, Singh and Holmes, 1996) or deep learning (Ruder, Ghaffari and Breslin, 2016). Similarly, many different feature sets can be chosen.

Feature Sets

In theory, almost anything that can be extracted from text can be used as a feature. Rudman (1998) states that more than a thousand different features have been proposed for this task, but the bulk of research has focused on three main types of features. Mosteller and Wallace (1964) and Binongo (2003) focused on words, as (at one remove) did De Morgan (1851), who focused on the distribution of the lengths of individual words. Other researchers such as Stamatatos (2009; 2013) have concentrated on characters, and specifically on character *n*-grams, clusters of *n* adjacent characters, such as the "sid" found in the word "inside". (Word *n*-grams, clusters of adjacent words, are also studied as a derivative of words themselves). Finally, some studies (Koppel and Schler, 2003; Zhai and Zobel, 2007) have investigated syntactic classes such as parts-of-speech or their derived *n*-grams. [While these three feature types account for the bulk of research attention, there are of course many others, including punctuation (Abbasi and Chen, 2008), layout (Zheng et al., 2006), and even color (Abassi and Chen, 2006)].

These methods share several advantages. They can generally be extracted with high efficiency, reliability, and accuracy. They do not require high levels of linguistic expertise to evaluate. There are even psycholinguistic reasons why this type of low-level feature may be good measure of writing style. Research (Bransford, Barclay and Franks, 1972) has shown that when human subjects listen to sentences, they do not pay much attention to the details of the specific words, especially common/function words, used. Subjects could not remember whether the sentence they had heard a moment ago was "Three turtles sat on a log and a fish swam under them" or "Three turtles sat on a log and a fish swam under it" – in either case, the listeners (presumably) had simply understood the meaning of the sentence, constructed a mental model describing the event, and

failed to appreciate fully that there are multiple ways to describe that single event.

Similarly, a classic brain teaser illustrates some of the issues with this kind of word. *Count the number of F's in the following sentence: FINISHED FILES ARE THE RESULT OF YEARS OF SCIENTIFIC STUDY COMBINED WITH THE EXPERIENCE OF YEARS.* Most people count three, missing the F's in the word "OF".

However, this very invisibility means that the conclusions drawn from this type of analysis are difficult to interpret in terms of traditional literary analysis. It's difficult to imagine a literary article discussing Doris Lessing's use of prepositions – instead, a typical comment would refer to "that epicist of the female experience, who with skepticism, fire, and visionary power has subjected a divided civilization to scrutiny" (Nobel Committee, 2007). As the title of a famous paper (Craig, 1999) put it, "if you can tell authors apart, have you learned anything about them?". Knowing that one author uses more instances of the letter "i" (McDonald et al., 2012) or fewer instances of "by" (Mosteller and Wallace, 1964) does not tell people anything most would find useful, or even informative. Even a relatively transparent inference – for example, a high percentage of definite articles indicates a relative preponderance of nouns – does not actually inform scholarship or understanding in any meaningful way.

What types of information would, then, truly inform? Word lengths (De Morgan, 1851) might actually be informative. They have long been considered to be a mark of intelligence and/or education; for example, the Flesch-Kincaid Reading Grade Level (Kincaid et al, 1975) is based on a combination of word length and sentence length; the routine use of long words probably indicates, at a minimum, that the author of the text has a large expressive vocabulary. At the same time, word (and sentence) length retain some of the objective accuracy and computational efficiency of other typical feature sets.

What other feature sets might provide this combination of computational and interpretational advantages? We propose that "concepts", the basic ideas expressed in writings, may provide clues to authorship. For example, to write convincingly of a city may require knowledge of that city. The Catholicism expressed by the fictitious *Father Brown* stories reflects Chesterton's own knowledge of (and belief in) Catholic precepts (Petersen, 1996). The detective novel *The Cuckoo's Calling* shows many conceptual quirks, such as detailed descriptions of women's clothing, magical feelings, giant men, oversized front teeth, and overbearing schoolteachers (Marsden, 2013; Vineyard, 2013). These concepts are encoded into the story at least in part in terms of the vocabulary used. By applying a semantic ontology of the words used in a work, we can determine not just the words a person writes but the concepts they express.

Materials and Methods

Our Corpus

One of the leading current mysteries is the true identity of the best-selling author Elena Ferrante (Tuzzi and Cortelazzo, 2018). While "her" influence is no doubt great (she has come close to winning the Strega prize twice; her novels have become films, and perhaps most significantly, she has been listed among the hundred most relevant works in 20th century Italian literature), little is known about her. She is presumed to be a female writer and equally presumed to be from the Neapolitan region, but (unlike the Rowling case) neither assumption has been confirmed (see Mikros, this volume). Instead, there has been a substantial outpouring of analysis identifying various potential authors as the real person behind this pseudonym.

In consequence of this literary mystery, Tuzzi and Cortelazzo (2018; see also Tuzzi and Cortelazzo, this volume) have collected a large corpus of novels from the past three decades. In addition to seven Ferrante novels, the corpus contains 143 other works (150 in total), written by forty different authors (including Ferrante). There are twelve non-Ferrante female authors, represented by forty-three books. The Neapolitan area is represented by ten non-Ferrante authors (thirty-nine books). This ten-million word corpus thus represents one of the largest-scale corpora ever assembled specifically for stylometry. Several stylometric scholars were invited to analyze this collection in the hopes of shedding light on Ferrante's mysterious identity.

This corpus thus provides an unusual (and valuable) opportunity to cross-compare several different techniques and methods. In addition to the more well-established methods of assessing authorship, the number of participants enables the comparison of new methods and feature sets. We therefore see this as an ideal opportunity to explore authorship attribution by concepts.

Our Feature Sets: A Conceptual Ontology

While the phrase "conceptual ontology" is relatively new (Google Books Ngram Viewer dates it only to 1965), the idea itself is much older. A thesaurus, for example, is simply a list of words organized into near-synonymous sets ("synsets") that may or may not be organized into larger hierarchies. For example, Roget's 1911 English Thesaurus (Roget, 1911) lists the following as expressing the idea of "variation" :

> variation; alteration, modification, moods and tenses; discrepance, discrepancy. divergency; deviation; aberration; innovation. vary; deviate; diverge; alternate, swerve. varied; modified; diversified.

Several key aspects of these synsets should be noted. First, as is common, the words themselves are not typically inflected (e.g, "variation" is included but not "variations"; "deviate" is included but not "deviating" or "deviated"). Second, words can be assigned to multiple synsets: the word "innovation" appears in the "variation" synset but also in the "difference at different times" synset and the "newness" synset. Third, these synsets can include phrases as well as words; for example, the "retrospective time" synset (Roget, 1911) includes phrases like "the good old days" or "ancient times". Fourth, these synsets themselves are often widely available (for example, by download), making them a practical resource for computational exploitation.

Fifth and most importantly, the assignment of words to synsets has typically been performed outside of any particular data set by domain experts. This stands in stark contrast to analysis techniques like topic modelling (Underwood, 2012). Like a thesaurus, topic modelling seeks to create sets of words that, collectively, describe a concept. However, these sets are defined probabilistically in terms of co-occurrence – two words are in the same set because when one appears in a document, the other probably does as well, and when one fails to appear, so, probably does the other. Computers infer these sets by statistical analysis of documents, but humans are still required to make sense of them – and often, topic word lists contain words that are related by something other than meaning. For example, the words "scarecrow", "tin" and "lion" might be linked, not by meaning, but by co-reference to a culturally salient artifact (*The Wonderful Wizard of Oz* and its sequels).

Once these synsets have been identified, it is relatively simple task for a computer to determine the number of instances of the words/phrases belonging to a particular synset that appear in a particular document. These can be expressed in relative terms as a token percentage that would vary from document to document, and also from document category (e.g., "all documents by Author A") to document category ("all documents by Author B"). Thus these synsets constitute a possible feature set for comparing authors using ordinary classification methods.

Our Method: Counting Significant Differences

For the classification experiment described in this paper, we experimented with a novel approach based on the conjecture generation software described in Juola (2009; 2010; 2012). In broad terms, this approach involves random slot-filling in a conjecture scheme, where each conjecture describes a simple but testable proposition from a large domain-specific vocabulary.

Among the earliest examples of this type of program was the Graffiti program (Fajtlowicz, 1988). This program used a large catalog of concepts (more

formally, graph-theoretic invariants) from graph theory, such as the number of leaves and the number of colors it would take to draw the graph with no two adjacent vertices sharing the same color. Graffiti would randomly pick two concepts, guess that the first is always at least as big as some expression involving the second, then test thousands or millions of graphs to see whether, in fact, this conjecture holds in all studied cases. (If it doesn't, mathematicians know the conjecture is false. If it does, then mathematicians might believe that it always holds and try to prove it. Graffiti has resulted in more than 100 publications over the past thirty years).

Our text analysis version, the Conjecturator, uses a similar random slot-filling approach to create statistically testable conjectures in large corpora. For example, "Do <words of a named category> appear more in <documents of a first type> than in <documents of a second type>?" For example, novels written by men might have fewer uses of scent-related adjectives than novels written by women. Analysis by the Conjecturator on a collection of Victorian novels confirmed this. Where does this come from? It could simply be a cultural artifact, but there is also the possibility, given that biologists have confirmed that women generally have better senses of smell than men, that there is a universal biological basis. This simple example demonstrates that this type of conjecture generation can not only demonstrate differences between groups, but also provide material to enrich future scholarship about these groups.

In this study, the document types were, of course, simply the novels by a given author. The named categories of words were thirty synsets extracted from an online Italian thesaurus, Sinonimi Master (available at http://www.homolaicus.com/linguaggi/sinonimi/index.htm). This thesaurus was scraped by hand to extract thirty different synsets. (Perhaps obviously, more synsets could be collected, possibly automatically, but these thirty suffice as a proof-of-concept). As an example, the synset associated with *abbandonato* includes:

> *abbandonato, arcaico, cadente, deserto, desolato, desueto, dimenticato, disabitato, escluso, incolto, incustodito, morto, negletto, obsoleto, recondito, remoto, selvaggio, spopolato, trascurato*

Calculating the type frequency of all words in this list yields an estimate of the frequency (also the salience and the importance) of that concept in any particular document. Averaging such estimates across all documents by a given writer in the corpus provides an estimate of the use of that concept to that particular writer. Finally, ordinary *t*-tests can tell us whether an observed difference between two writers is "significant" (in the statistical sense), and possibly therefore whether it is suggestive of anything interesting in the literary analysis sense.

Some samples of our findings include (p-values are the results of the t-tests):

- The word group *abbandonato* does not vary between Morazzoni and Mazzucco ($p \sim 0.1393$)
- The word group *abbandonato* does not vary between Vinci and Scarpa ($p \sim 0.3212$)
- The word group *abbandonato* appears more in Veronesi than in Ferrante ($p \sim 0.9810$)

This provides evidence of a single aspect of difference between Veronesi and Ferrante (and thus shows that Veronesi is probably not Ferrante), but also provides a specific research question for literary scholars: Why and how does Veronesi use the concept of *abbandonato* (abandoned, derelict, forsaken) so much?

Results

Our computers generated and tested 10,000 random conjectures. As these conjectures were random, not all involved Ferrante, but every author was compared to Ferrante between 10 and 25 times, measuring the use of ten to twenty-five different concepts.

Of the thirty-nine distractor authors, only two (Murgia and Starnone) had no significant differences in concept usage when compared with Ferrante.

Discussion

The previous section showed that, as far as this preliminary analysis is concerned, both Murgia and Starnone use the same set of concepts as Ferrante. From this, our preliminary conclusion is, naturally, that Murgia and Starnone are (jointly) the most likely members of this author set to be the author behind Ferrante's ghost-writing.

Of course, the standard warnings apply. This framework and conclusion implicitly assumes a so-called closed-class problem, where the author must be one of the writers in the set. If we allow none-of-the-above as a possible response, then all that has been shown is that Ferrante is someone who writes like Starnone and Murgia.

Because of the scale of this preliminary experiment, it is not possible to say with confidence whether Murgia or Starnone is the most probable author among the distractor set. This can be easily addressed simply by using more synsets. Similarly, the efficiency of the current setup for addressing the particular question of interest ("Who is Elena Ferrante?") is questionable, as the computer spent as much or more time addressing conjectures about differences between

distractor authors themselves as it did with analyzing Ferrante's writing style. A more systematic exploration could, for example, compare a thousand extracted synsets between Ferrante and each individual candidate author, providing up to a thousand differences; developing this program would be a simple task for an appropriately skilled programmer, but has not been done.

There are similarly a number of issues related to the handling of the synsets themselves. Because words can be polysemous, synsets can be ambiguous – any individual word token will be counted as part of all synsets its type is a member of. Similarly, the system does not even attempt to address inflectional morphology – "*abbandonato*" can be a past participle verb, but the computer as currently programmed will not recognize or tabulate present tense forms such as "*abbandoniamo*" or simple future forms such as "*abbandonerà*". The program does not distinguish between verb and adjective use, but also will not recognize "*abbandonata*", "*abbandonati*", and "*abbandonate*".

We are nevertheless heartened by this result as a proof-of-concept. The results presented above show, first, that conceptual usage can be measured in a computationally tractable way. Second, conceptual usage can be a feature that distinguishes one writer from another (and by extension, can probably distinguish one social group from another). Third, the process of evaluating concept use will automatically generate human-interpretable data describing the differences between authors in meaningful terms. In Craig's words, if we can tell two authors apart, we have learned something very meaningful about them – we have learned the ideas they try to express.

References

Abbasi, A. and Chen, H. (2006). Visualizing authorship for identification. In *Proceedings of the 4th IEEE Symposium on Intelligence and Security Informatics*, San Diego, 60-71.

Abbasi, A. and Chen, H. (2008). Writeprints: A Stylometric Approach to Identity-level Identification and Similarity Detection in Cyberspace, *ACM – Transactions on Information Systems*, 26(2), 7.

de Morgan, A. (1851). Letter to Rev. Heald 18/08/1851. In S. E. de Morgan (ed.), *Memoirs of Augustus de Morgan by his wife Sophia Elizabeth de Morgan with Selections from his Letters*. London: Longman's Green and Co., 1882.

De Vel, O., Anderson, A., Corney, M. and Mohay, G. (2001). Mining e-mail content for author identification forensics, ACM SIGMOD Rec. 30(4), 55-64.

Binongo, J. N. G. (2003). Who wrote the 15th book of Oz? An application of multivariate analysis to authorship attribution, *Chance* 16(2), 9-17.

Bransford, J.D., Barclay, J.R. and Franks, J.J. (1972). Sentence memory: A constructive versus interpretive approach, *Cognitive psychology*, 3(2), 193-209.

Craig, H. (1999). Authorial attribution and computational stylistics: If you can tell authors apart, have you learned anything about them?, *Literary and Linguistic Computing*, 14(1), 103-113.

Diederich, J., Kindermann, J., Leopold, E. and Paass, G. (2003). Authorship attribution with support vector machines, *Applied intelligence* 19(1-2), 109-123.

Fajtlowicz, S. (1988). On conjectures of *Graffiti*, *Discrete Mathematics*, 38, 113-118.

Friedman, W.F. and Friedman E. S (1957). *The Shakespearean Ciphers Examined: an analysis of cryptographic systems used as evidence that some author other than William Shakespeare wrote the plays commonly attributed to him*, Cambridge: Cambridge University Press.

Joachims, T. (2002). *Learning to Classify Text Using Support Vector Machines*. Norwell: Kluwer.

Juola, P. (2008). Authorship attribution, *Foundations and Trends® in Information Retrieval*, 1(3), 233-334.

Juola, P. (2009). Mapping genre space via random conjectures. In *Chicago Colloquium on Digital Humanities and Computer Science* (DHCS), Chicago, November 2009.

Juola, P. (2010). Distant Reading and Mapping Genre Space via Conjecturebased Distance Measures. Digital Humanities 2010, London.

Juola, P. (2012). Automatic Analysis of Gender Variation in Language Via Conjecture Generation. Research Foundations for Understanding Books and Reading in the Digital Age: E/Merging Reading, Writing, and Research Practices. INKE 2012 BirdsofaFeather Gathering, Havana.

Juola, P. (2015). The Rowling Case: A Proposed Standard Protocol for Authorship Questions. *Digital Scholarship in the Humanities*, 30(1), i100-i113.

Kincaid, J.P., Fishburne, R.P., Rogers, R.L. and Chissom, B.S. (1975). *Derivation of new readability formulas (automated readability index, fog count, and Flesch reading ease formula) for Navy enlisted personnel. Research Branch Report 8–75*. Chief of Naval Technical Training: Naval Air Station Memphis.

Koppel, M. and Schler J. (2003). Exploiting stylistic idiosyncrasies for authorship attribution. In *Proceedings of IJCAI'03 Workshop on Computational Approaches to Style Analysis and Synthesis*, 69, 72-80.

Koppel, M., Schler, J. and Argamon, S. (2009). Computational methods in authorship attribution, *Journal of the Association for Information Science and Technology*, 60(1), 9-26.

McDonald, A.W.E., Afroz, S., Caliskan, A., Stolerman, A. and Greenstadt, R. (2012). Use fewer instances of the letter "i": Toward writing style anonymi-

zation. In *International Symposium on Privacy Enhancing Technologies Symposium*, 299-318. Berlin-Heidelberg: Springer.

Marsden, S. (2013). The Cuckoo's Calling: publishers' embarrassment at turning down JK Rowling detective novel, *The Telegraph*, 14 July 2013 (http://www.telegraph.co.uk/culture/books/10178960/The-Cuckoos-Calling-publishers-embarrassment-at-turning-down-JK-Rowling-detective-novel.html).

Mosteller, F. and Wallace, D.L. (1964). *Inference and disputed authorship: The Federalist*. Reading: Addison-Wesley.

"The Nobel Prize in Literature 2007 – Press Release". *Nobelprize.org*. Nobel Media AB 2014. Web. 5 Feb 2018 (http://www.nobelprize.org/nobel_prizes/literature/laureates/2007/press.html).

Noecker Jr, J. and Juola P. (2009). Cosine Distance Nearest-Neighbor Classification for Authorship Attribution, *Proceedings of Digital Humanities 2009*, College Park, Maryland.

Petersen, J. (1996, ed.). *Father Brown of The Church of Rome: Selected Mystery Stories by G.K. Chesterton*. San Francisco: Ignatius Press.

Porter, S.E. (1995). Pauline Authorship and the Pastoral Epistles: Implications for Canon, *Bulletin for Biblical Research* 5, 105-123.

Quinlan, J.R. (1993). *C4.5: Programs for Machine Learning*. San Mateo: Morgan Kauffman.

Roget, P.M. (1911). Roget's Thesaurus No. 2. (1911). Project Gutenberg. 5 Feb 2018. < http://www.gutenberg.org/cache/epub/22/pg22.txt>

Ruder, S., Ghaffari, P. and Breslin J.G. (2016). Character-level and multi-channel convolutional neural networks for large-scale authorship attribution. *arXiv preprint arXiv:1609.06686*.

Rudman, J. (1997) The state of authorship attribution studies. Some problems and solutions, *Computers and the Humanities*, 31(4), 351-365.

Stamatatos, E. (2009). A survey of modern authorship attribution methods, *Journal of the Association for Information Science and Technology* 60(3), 538-556.

Stamatatos, E. (2013). On the robustness of authorship attribution based on character n-gram features, *Journal of Law and Policy*, 21(2), 420–440.

Tuzzi, A. and Cortelazzo M.A. (2018). What is Elena Ferrante? A comparative analysis of a secretive bestselling Italian writer, *Digital Scholarship in the Humanities* (online first 19 January 2018 fqx066, https://doi.org/10.1093/llc/fqx066).

Tweedie, F.J., Singh, S., and Holmes D.I. (1996). Neural network applications in stylometry: The Federalist Papers, *Computers and the Humanities* 30(1), 1-10.

Underwood, T. (2012). Topic modeling made just simple enough. Blog post of April 7 2012 (https://tedunderwood.com/2012/04/07/topic-modeling-made-just-simple-enough/ Accessed 7 February 2018).

Vineyard, J. (2018). 10 Harry Potter Hallmarks Found in J.K. Rowling's The Cuckoo's Calling. Vulture.com. 5 February 2018 (http://www.vulture.com/2013/07/jk-rowling-cuckoos-calling-harry-potter-links.html).

Zhao, Y. and Zobel J. (2007). Searching with style: Authorship attribution in classic literature. In *Proceedings of the thirtieth Australasian conference on Computer science-Volume 62*, Australian Computer Society, Inc., 59-68.

Zheng, R., Li, J., Chen, H., and Huang, Z. A. (2006). A Framework for Authorship Identification of Online Messages: Writing-Style Features and Classification Techniques, *Journal of the American Society for Information Science and Technology*, 57(3), 378-393.

Data-Compression Approach to Authorship Attribution

Margherita Lalli[1], Francesca Tria[1] and Vittorio Loreto[2,1,3]
[1] *Sapienza Univ. of Rome, Physics Dept., Piazzale Aldo Moro 2, 00185 Rome, Italy*
[2] *SONY Computer Science Laboratories, 6, rue Amyot, 75005, Paris, France*
[3] *Complexity Science Hub Vienna, Josefstädter Strasse 39, 1080 Vienna, Austria*

Abstract

Authorship attribution is a fascinating field at the crossroad between linguistics and information science. Its relevance goes much beyond the specific predictions that different tools can make about authors whose identity is uncertain or hidden behind known "noms de plume". Correctly spotting the unknown author of a text is far from reflecting a "keyhole" attitude, representing instead the tip of an iceberg whose main body is made of solid tools and algorithms able to extract syntactic, possibly semantic, information out of generic strings of characters. Here we follow a data-compression approach to authorship attribution through which we define a notion of similarity between generic strings of characters (in particular literary texts). We start by assessing the overall performance of our set of tools in performing authorship attribution both on the wide corpus adopted in this volume and on an extended corpus. We then concentrate on the well-known "affaire Ferrante" (originally treated by some of us back in 2006[1]), confirming and strengthening our original claim that, within the corpus considered, Domenico Starnone is the most likely author behind Elena Ferrante. We stress again that, despite the strong hints pointing to Starnone, we cannot rule out the possibility that Ferrante's signature could hide another author (or several authors) not included in the corpus. Specific analyses are still in order to shed light on this last point.

[1] L. Galella, *Ferrante è Starnone. Parola di computer*, L'Unità, 23 November 2006.

Introduction

In nature, many systems and phenomena are often represented in terms of sequences or strings of characters. In experimental investigations of physical processes, for instance, one typically has access to the system only through a measuring device which produces a time record of a certain observable, i.e. a sequence of data. On the other hand, other systems are intrinsically described by string of characters, e.g. DNA and protein sequences, written texts. When analyzing a string of characters, the main aim is to extract the information it provides. For a DNA sequence this would correspond, for instance, to the identification of the sub-sequences codifying the genes and their specific functions. For time series (Badii and Politi, 1997), one could be interested in the extraction of specific features or trends. On the other hand, for a written text one is interested in questions like recognizing the language in which the text is written, the subject treated or its author (see Stamatatos, 2009) for a relatively recent review of authorship attribution methods). Key to this end is the definition of suitable quantities to quantify the similarity/remoteness of two strings of characters, and more specifically between two texts. With this aim in mind, it is rather natural to approach the problem from the point of view of Information Theory (IT). Born in the context of electric communications, IT theory has acquired, since the seminal paper of Shannon (Shannon, 1948), a leading role in many other fields as computer science, cryptography, biology and physics (Zurek, 1990). In this context, the word information acquires a very precise meaning, namely that of the entropy of the string, a measure of the surprise the source emitting the sequences can reserve to us.

It is important to stress that IT deals with ensembles of sequences emitted by an ergodic source, while one is typically forced to treat a single sequence. In this spirit, an appropriate concept is that of algorithmic complexity (AC) (Kolmogorov, 1965; Chaitin, 1966; 1990; Solomonoff, 1964). The AC, also known as Kolmogorov complexity, of a string of characters is given by the length (in bits) of the smallest program which produces as output the string and stops afterwards. A string is said to be complex if its complexity is proportional to its length. This definition is really abstract, in particular it is impossible, even in principle, to find such a program (Li and Vitányi, 1997). Despite the impossibility to compute the AC of a sequence, there are algorithms explicitly conceived to give a good approximation to it (Li and Vitányi, 1997). In particular, since the AC of a string fixes the minimum number of bits one should use to reproduce it (optimal coding), it is intuitive that a typical zipper, besides trying to reduce the space occupied on a memory storage device, can be considered as a meter for the Algorithmic Complexity of a generic string. The better will be the compression algorithm, the closer will be the length of the zipped file to the optimal

coding limit and the better will be the estimate of the AC provided by the zipper. It is well known that compression algorithms represent a powerful tool for the estimation of the AC or more sophisticated measures of complexity complexity (Ziv and Merhav, 1993; Milosavljevíc, 1995; Farach *et al.*, 1995). Several applications have been drawn in several fields (Verdú, 1998) from dynamical systems theory (the connections between IT and dynamical systems theory are very strong and go back all the way to the work of Kolmogorov and Sinai; for a recent overview see Lind and Marcus, 1995; Benci *et al.*, 2002; Boffetta *et al.*, 2002) to linguistics (an incomplete list would include: Bell, Cleary and Witten, 1990; Puglisi *et al.*, 2003; Teahan, 2000; Juola, 1998; El-Yaniv, Fine and Tishby, 1997; Thaper, 2001; Kukushkina, Polikarpov and Khmelev, 2000; Benedetto, Caglioti and Loreto, 2002) and genetics (Li *et al.* 2001; Grumbach and Tahi, 1994; Loewenstern, Hirsh, Yianilos and Noordewieret, 1995). Some of us have recently proposed a method (Benedetto, Caglioti and Loreto, 2002) for context recognition and context classification of strings of characters or other equivalent coded information. The remoteness of a sequence B from a sequence A was estimated by zipping a sequence A + B obtained by appending the sequence B after the sequence A and using the gzip compressor – whose core is provided by the Lempel–Ziv 77 (LZ77) algorithm (Lempel and Ziv, 1977). This approach was adopted for authorship attribution and, defining a suitable distance between sequences, for reconstructing phylogenetic language trees. The idea of appending two files and zip the resulting file in order to measure the remoteness between them had been previously proposed by Loewenstern *et al.* (1995) (using zdiff routines) who applied it to the analysis of DNA sequences, and by Khmelev and coworkers (2000) who applied the method to authorship attribution. Similar methods have been proposed by Juola (1998), Teahan (2000) and Thaper (2001). Later on, the method proposed by Benedetto, Caglioti and Loreto (2002) was further refined (Baronchelli *et al.*, 2005) thanks to the insights reported in (Puglisi *et al.*, 2003), where it was investigated how a compression algorithm optimizes its features at the interface between two different sequences A and B while zipping the sequence A + B obtained by simply appending B after A. It was shown the existence of a scaling function (the "learning function") which rules the way in which the compression algorithm learns a sequence B after having compressed a sequence A. In particular, it turns out that there exists a cross-over length for the sequence B, which depends on the relative entropy between A and B, below which the compression algorithm does not learn the sequence B (measuring in this way the cross-entropy between A and B) and above which it starts learning B, i.e. optimizing the compression using the specific features of B. With these insights, a new compression scheme new compression scheme (Puglisi *et al.*, 2003) was devised to measure the remoteness between two texts. In a nutshell,

the new scheme forces the algorithm to learn only from sequence A when compressing B, thus avoiding the drawbacks of the "learning function" discussed above.

The outline of the paper is the following. In the next section we discuss in details the compression algorithm we adopt in this paper, along as the specific entropic quantities we shall be measuring and adopting for our authorship attribution task. In the Section 3 we briefly discuss the corpus adopted in this paper and the preprocessing performed. Section 4 presents the results organized in two main lines: the overall reliability of the method when tested on the whole corpus and the specific problem of identifying the most likely *plume* behind Elena Ferrante. We finally draw some conclusions and discuss open challenges.

Data-compression based tools

In this section we describe the data-compression scheme we adopt in this paper for the authorship attribution task. The method was first introduced in (Baronchelli *et al.*, 2005) and we refer to it for its details. Here we remind only the main definitions and the functioning principles. The method is based on the LZ77 compression algorithm (Lempel and Ziv, 1977), which, roughly speaking, achieves the compression of a file exploiting the presence of repeated subsequences. More in details (see Figure 1) let $x = x_1, \ldots, x_N$ be the sequence to be zipped, where x_i represents a generic character of a sequence's alphabet. The LZ77 algorithm finds duplicated strings in the input data. The second occurrence of a string is replaced by a pointer to the previous string given by two numbers: a distance, representing how far back into the window the sequence starts, and a length, representing the number of characters for which the sequence is identical. More specifically, the algorithm proceeds sequentially along the sequence. Let us suppose that the first n characters have been codified. Then the zipper looks for the largest integer m such that the string $x_{n+1}, \ldots, x_{n+m}$ already appeared in $x_1, \ldots, x_n$. Then it codifies the string found with a two-number code composed by: the distance between the two strings and the length m of the string found. If the zipper does not find any match then it codifies the first character to be zipped, x_{n+1}, with its name. This eventuality happens for instance when codifying the first characters of the sequence, but this event becomes very infrequent as the zipping procedure goes on.

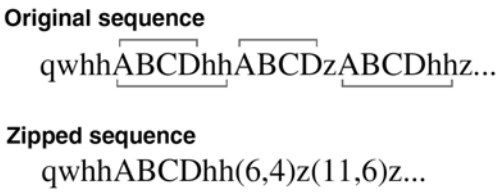

Figure 1. Scheme of the LZ77 algorithm: the LZ77 algorithm works sequentially and at a generic step looks in the look-ahead buffer for substrings already encountered in the buffer already scanned. These substrings are replaced by a pointer (d, n) where d is the distance of the previous occurrence of the same substring and n is its length. Only strings longer than two characters are replaced in the example.

The way in which we exploit this compression scheme to measure the remoteness between two texts involves the notion of cross-entropy. Let us consider two stationary zero-memory sources A and B emitting sequences of 0 and 1: A emits a 0 with probability p and 1 with probability 1 – p while B emits 0 with probability q and 1 with probability 1–q. A compression algorithm like LZ77 applied to a sequence emitted by A will be asymptotically (i.e., in the limit of an available infinite sequence) able to encode the sequence almost optimally (Wyner and Ziv, 1994), i.e., coding on average every character with $[-p \log_2 p - (1-p) \log_2 (1-p)]$ bits (the Shannon entropy of the source). This optimal coding will not be the optimal one for the sequence emitted by B. In particular the entropy per character of the sequence emitted by B in the coding optimal for A (i.e., the cross-entropy per character) will be $[C(p|q) = -q \log_2 p - (1-q) \log_2 (1-p)]$ while the entropy per character of the sequence emitted by B in its optimal coding is $[C(q|q) = -q \log_2 q - (1-q) \log_2 (1-q)]$. Notice C(p|q) is always greater than C(q|q) and C(p|q) approaches C(q|q) from above when q approaches p. More generally, the cross-entropy between two sources (represented by two probability distributions) measures the number of bits needed to encode messages from the one of the sources when using the optimal code for the other source. A linguistic example will help to clarify the situation: transmitting an Italian text with a Morse code optimized for English will result in the need for transmitting an extra number of bits with respect to another coding optimized for Italian. The cross entropy between Italian and English in this case will be given by the number of bits per character needed to code a message in Italian using the Morse code optimized for English. Closer are Italian and English, smaller will be the corresponding cross-entropy between them.

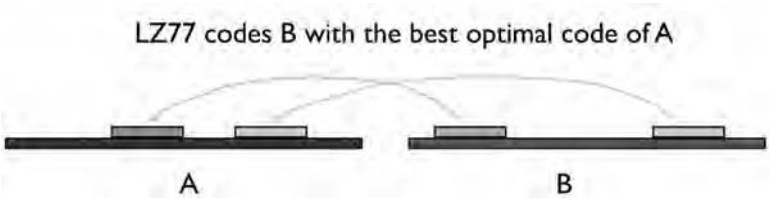

Figure 2. Scheme of the LZ77 algorithm when comparing two different texts. Instead of concatenating the two texts and zipping them sequentially (which would imply that the compressor first learns A, then for a "while" (see (Benedetto et al., 2002)) measures the cross-entropy between B and A and then learns B), one always scans B. In the refined scheme the compressor does not zip the B part but simply 'reads' it with the (almost) optimal coding of part A. In this case we start reading sequentially file B and search in the look-ahead buffer of B for the longest subsequence that already occurred only in the A part. This means that we do not allow for searching matches inside B itself. As in the usual LZ77, every matching found is replaced with a pointer indicating where, in A, the matching subsequence appears and its length.

The main idea to exploit the notion of cross-entropy to measure the remoteness between two texts is thus the following. Given two texts A and B (emitted by two different "sources[2]", i.e., two different authors) one can use a LZ77-like compression scheme to encode B given the best code of A. In practice this implies scanning the B text and looking for matches only in the A text. The procedure is illustrated in Figure 2.

This method allows us to measure (or at least to estimate) the cross-entropy of B with respect to A, i.e., C(A|B). If the two texts are very similar the number of bits needed to encode B through the knowledge of A, i.e., C(A|B), will be correspondingly small. On the other hand, if B and A are very different the knowledge of A will not help encoding B and C(A|B) will be correspondingly large. We refer to (Baronchelli et al., 2005) for a more detailed description of the overall method. Here it is enough to say that given two texts A and B our data-compression oriented scheme allows to estimate their cross-entropy C(A|B) and we shall use C(A|B) as measure of remoteness between A and B.

An interested reader, may find in (Juola and Baayen, 2005) an alternative

[2] Notice that, in the framework of authorship attribution, it is difficult to speak about sources. Strictly speaking, authors cannot be considered sources because we don't know whether a probability distribution through which they emit symbols is well defined in the first place: is it stationary? is it changing over time? Notice that the same caveat applies to the very definition of a Language. Is Language representing a well defined source? Which English? Written by whom? Despite these difficulties here we shall keep considering individual Languages and texts as emitted by specific sources (speakers or writers in that Language or Authors of given texts).

attribution tool based on cross-entropy measures.

Given the above-defined method, we can now introduce a general scheme for authorship attribution, that we will later refine. Suppose that we are interested in the identification of the putative author of a given text X. The procedure we use considers a collection (a corpus), as large as possible, of texts from known authors, each denoted Y_i; where Y denotes the author and *i* denotes the *i*-th text of the author Y. Given this corpus of texts, that we wish to be as large as possible, we compute all the cross-entropies, $C(Y_i|X)$ of the text X with all the texts Y_i of the corpus and we look for the text Y_i^* for which the cross-entropy with the text X is minimum. The author of the text Y_i^* is candidate to be the author of the text X. In practice, this scheme can be refined, by considering further information beside the one given by Y_i^*. In the following sections we shall provide the reader with more details about the specific algorithms to perform the attribution. Here it is enough to say that our data-compression scheme allows to give a measure of remoteness between pairs of texts and this will be basis for the attribution procedure.

Before concluding this section, several remarks are in order concerning our minimum cross-entropy method used to perform authorship attribution. Our criterion has been that of saying that the text X should be attributed to a given author if another work of this author is the closest (in the cross-entropy ranking) to X. It can happen that this is not a good criterion, both for an intrinsic reason, for instance the large variety of features that can be present in the production of an author, and for an extrinsic one, due for instance to an under-sampling of the production of that author. We shall come back to those points in the following.

A further remark concerns the fact that our results for authorship attribution could only provide some hints about the real paternity of a text. One cannot, in fact, ever be sure that the reference corpus contains at least one text by the unknown author. If this is not the case we can only say that some works of a given author resemble the unknown text. This remark is particularly relevant for the *affaire* Elena Ferrante and we shall come back on this point later on.

The corpus adopted

We refer to two sets of data: the corpus of 150 Italian texts described in first sections of this book and an extended version of it which includes in addition 29 texts written by authors already present in the original corpus. In most of the cases we shall report our measures for both corpus or, otherwise, it will be specified and justified. In Table 1 we enumerated the added texts. This supplement does not affect general results except for an improvement in the performance of

the method as the amount of information for each reference author increases. This improvement follows from the nature of the method, which allows us to identify the author of a text comparing it with each text with known authorship in the corpus: the higher is the number of works of a certain author, the higher is the probability to correctly recognize him in the unknown text. Imagine in fact that Z is the author of the unknown text X and in the reference corpus we only have texts of Z belonging to a very different period with respect to the one in which she/he wrote X. It is likely in this case that Y_i^* will be a text of an author Y, different from Z. On the other hand, if we would have the entire production of Z in the corpus, and still the method fails in correctly attributing X, we could think of at least two different reasons: (i) the method is not accurate enough; (ii) the author is so eclectic that different authors are more similar to her/his text than herself.

Following the scheme of information theory outlined in the previous sections, we consider a text as a symbolic sequence of characters: the units analyzed are alphabetical characters (where upper and lowercase are considered different), punctuation marks, numbers and blank spaces. Therefore, in preprocessing the data we removed counting of pages and chapters, together with their titles, and replaced line breaks and multiple spaces with single space.

Further, since the data compression scheme we adopt (i.e., LZ77) is strongly sensitive to the length of the sequences considered (a scheme as LZ77 continuously learns to better compress while it processes the sequence of characters), and in particular of the texts used as references, we took all the reference texts of equal size. To this end we segment each text in the corpus into fragments of equal size and use these fragments as reference texts. In our case, we chose the size of the fragments to be 29566 bytes (characters), i.e., the size of the shortest text within Elena Ferrante's work (*La Frantumaglia*). Since the method is in principle more accurate longer are the texts, we also tested the robustness of our results against different sizes of the fragments.

Table 1. A record of the additional texts included in the extended corpus in respect to original corpus.

Author	**Title**	**Publication Year**
De Luca	Non ora non qui	1989
De Luca	Sulla traccia di Nives	2005
De Luca	E disse	2011
De Luca	Il torto del soldato	2012
De Luca	La doppia vita dei numeri	2012

De Luca	La parola contraria	2015
De Silva	Voglio guardare	2002
De Silva	Mancarsi	2013
De Silva	Terapia di coppia per giovani amanti	2015
Ferrante	La frantumaglia	2003
Maraini	Un clandestino a bordo	1993
Maraini	Bagheria	1993
Maraini	Dolce per sé	1997
Maraini	La nave per Kobe	2001
Maraini	Passi affrettati	2007
Maraini	La ragazza di Via Maqueda	2009
Maraini	L'amore rubato	2012
Maraini	Chiara di Assisi. Elogio della disobbedienza	2013
Maraini	La bambina e il sognatore	2015
Mazzucco	Il bassotto e la regina	2012
Mazzucco	Limbo	2012
Mazzucco	Sei come sei	2013
Mazzucco	Io sono con te. Storia di Brigitte	2016
Ramondino	Althénopis	1981
Starnone	Labilità	2005
Starnone	Condom Butterfly	2008
Starnone	Spavento	2009
Starnone	Fare scene. Una storia di cinema	2010
Tamaro	Per sempre	2011

Results

The overall procedure and robustness

In this section we investigate the performance of the method in the authorship attribution task for every text in the corpus. More precisely, each text X of the corpus will be considered in turn as the text with unknown authorship, to be attributed by means of our method and the information given by the whole corpus but X. Since, as we mentioned in the previous section, after the preprocessing phase our sample is composed by fragments of texts, we have to set a procedure to attribute the whole text out of cross-entropy measures between couples of fragments. In the following we propose different strategies and give

results for each of them. The most accurate technique will be used in the attribution of Elena Ferrante's work.

Let us then consider our "unknown" text X and a reference text of the corpus Y_i, written by the author Y. Let us assume that X is segmented in 3 fragments (X^1,X^2,X^3) and Y_i in two (Y_i^1,Y_i^2). We can thus consider six different coupling between the fragments of X and those of Y_i, and the corresponding cross-entropy values:

$$\begin{array}{lll} C(Y_i^1|X^1) & C(Y_i^1|X^2) & C(Y_i^1|X^3) \\ C(Y_i^2|X^1) & C(Y_i^2|X^2) & C(Y_i^2|X^3) \end{array}$$

More in general, if the text X and the text Y_i are composed respectively of n and m fragments, we will have $n \times m$ cross-entropy values. We can then use the following procedures to attribute X.

Average on fragments cross-entropies

A measure of the remoteness between X and Y_i can be obtained with an arithmetic mean of all the pairwise cross-entropies between fragments:

$$C(Y_i|X) = \frac{1}{nm} \sum_{j,k}^{n,m} C(Y_i^j|X^k)$$

We then use two different criteria to attribute the text X:

(i) *First-nearest-neighbor approach:* we simply attribute X to the author of text Y_i^* with the lowest $C(Y_i|X)$.

(ii) *Weighted-profile approach:* we here exploit further information than the minimum value of the cross-entropy between the texts X and Y_i (averaged over fragments, as defined above). In particular, we wish to use the information coming from all the texts of a given author in the corpus, trying at the same time to reduce noise. To do that, we rank all the texts of a given author according to their cross-entropy with X in ascending order (the first in the rank is the one with lower cross-entropy with X). We then construct a weighted average for each author, that we use as a measure of distance[3] between the text X and the considered author (say Y):

$$D(Y|X) = \frac{\sum_{i=1}^{l} \frac{1}{r_i} C(Y_i|X)}{\sum_{i=1}^{l} \frac{1}{r_i}}$$

[3] We name it here *D* and from now on we may spell it *distance* but it is fair to outline that it is a pseudodistance since, as the cross-entropy, it does not satisfy the triangular inequality nor it is symmetric.

Where r_i is the relative rank of text Y_i, i.e. the rank of Y_i in a ranking where only texts of the author Y are present. The text X is attributed to the author of the corpus at minimum distance, that is the author Y for which D (Y|X) is minimum.

Majority-rule approach:

In this case we consider each fragment as carrying part of the information about the authorship of the entire text. Let the text X be composed of *n* fragments, we first attribute each fragment to an author, then (in a pure democratic approach, assuming we don't have any a priori information about the higher reliability of one fragment with respect to the other), we attribute the whole text X to the author to which the majority of the fragments point.

The attribution of a single fragment can be done in different ways, as discussed above for the entire text. In particular, we will use:

(i) *A first-nearest-neighbor approach on fragments*: the fragment X^k is attributed to the author Y for which the cross-entropy $C(Y_i^j|X^k)$ is minimum, for some fragment *j* of some text *i*.

(ii) *A weighted-profile approach on fragments*: we compute, for each fragment X^k and each author Y in the corpus, the (pseudo)distance:

$$D(Y|X^k) = \frac{\sum_s^{N_y} \frac{1}{r_s} C(Y^s|X^k)}{\sum_{s=1}^{N_y} \frac{1}{r_s}}$$

Where $N_y = n \times m$ is the total number of fragments of author Y and r_s is the relative rank of the fragment Y^s in a ranking where only fragments of the author Y are present. The fragment X^k is then attributed to the author with minimum $D(Y|X^k)$.

In the majority rule approach, all the other fragments of the same text as X^k are removed from the corpus when attributing X^k. Note also that in the majority rule approach a situation of parity can occur, in which the same maximum number of fragments is attributed to different authors. In our case this situation appears in only three cases, in the attribution of Nori, Scarpa and Valerio. In those cases, when one of the author is the correct one, we count a success score of 1/(# of attributed authors). For instance, the novel *Fermati un minuto a salutare* of Valerio has 7 fragments, of which 1 attributed to Faletti, 1 to Mazzucco, 1 to Lagioia, 2 attributed to Giordano and 2 to herself: we count then a success of ½ for that text.

In Table 2 we report cumulative results for the performance of the different methods we discussed, for different sizes of the fragments in which the texts are decomposed. It is evident that smaller fragments are more vulnerable to noise, and from now on we will refer on measures obtained from fragments 29566 bytes long.

Table 2. Here we show the percentage success rate for each attribution scheme discussed in the text, from fragments of length 29566, 20000 and 10000 bytes. In order of appearance, signatures on the left refer to: 1NN=*First-Nearest-Neighbor approach*, WP=*Weighted-Profile approach*, MR = *Majority-Rule when using a first-nearest-neighbor approach on fragments*, WMR = *Majority-Rule when using a weighted-profile approach on fragments*, WMR_30 = *Majority-Rule when using a weighted-profile approach on first 30 fragments in ranking*. WMR_90 = *Majority-Rule when using a weighted-profile approach on first 90 fragments in ranking.*

	Original corpus (%)			Extended corpus (%)		
	29566 bytes	*20000 bytes*	*10000 bytes*	*29566 bytes*	*20000 bytes*	*10000 bytes*
1NN	78	75	63	80	78	68
WP	82	74	66	82	79	73
MR	86	84	81	88	87	84
WMR	85	81	75	87	84	80
WMR_30	85	84	82	87	86	84
WMR_90	**87**	84	80	**89**	87	84

In Table 3 we list the success rates explicitly for every author in the corpus. When attributing the texts of Starnone, we excluded from the corpus all the texts of Elena Ferrante, and viceversa, when attributing the texts of Elena Ferrante, we excluded from the corpus all the texts of Starnone. We stress again that here we aim at accessing the ability of our method to perform authorship attribution, and since the attribution of a text of Starnone to Ferrante can be due both of a failure of our method, or to the fact that Starnone is indeed the author of the texts signed as Ferrante, we have to exclude this ambiguity a priori (in the next section we will face the problem of the attribution). In the case of the texts of Ferrante, it is interesting to access its self-consistence, that is to know if our method attributes all her texts to herself (when Starnone is excluded).

From Table 3 we see that the overall performance of 89% is not evenly distributed among the authors. In particular, for most of the authors our methods correctly attribute 100% of the texts.

However, there are few writers for which our methods largely fail. This can be due to several reasons. In the cases of Parrella and Vinci, for instance, the texts participating the corpus are only two, segmented, respectively, in 6 and 10 subtexts 29566 bytes long. This poor sample affects the probability of a text to be attributed correctly, as discussed above, and can be the cause of the failure of the method. The question of minimal sample size in the general context of attribution and classification of texts is still open, although largely surveyed (for a recent instance see Eder, 2017).

A different reason can be thought for the incorrect attribution of the text *Memorie di una ladra* of Dacia Maraini, whose anomalous nature can be ascribable to the wide chronological deviation from other works of the writer.

A separate issue deserves finally Tiziano Scarpa, whose sample, although copious, seems to be composed exclusively from anomalies: none of his texts is attributed to him. This could mark either some sort of eclecticism or a lack of originality or, still, none of them.

There are then intermediate cases, in which authors are partially self-consistent. An interesting issue concerning these situations relates to the margin of uncertainty connected with the attribution to any author. The question is: given a certain choice in the attribution of a text, for example in a majority rule scheme, what are the conditions preluding to that choice? The identified author to be, was it self-evident or controversial? Giving an answer means gaining a quantitative contribution to the measure of complexity of an author.

Table 3. Percentage success rate per author, in each attribution scheme discussed in the text, from fragments of largest size, i.e. 29566 bytes.

	Original corpus (%)					**Extended corpus (%)**				
	1NN	WP	MR	WMR	WMR_90	1NN	WP	MR	WMR	WMR_90
Starnone	90	100	95	100	93	100	100	100	100	100
De Luca	100	86	100	86	100	100	80	100	90	100
Carofiglio	100	100	100	100	100	100	100	100	100	100
Mazzucco	100	100	100	100	100	89	89	100	100	100
De Silva	100	100	100	100	100	100	100	100	94	100
Ferrante	100	100	100	100	100	88	88	100	100	100
Piccolo	86	86	100	90	100	86	87	100	100	100
Tamaro	100	100	100	100	100	100	100	100	100	100
Faletti	100	100	100	100	100	100	100	100	100	100
Mazzantini	75	100	100	100	100	75	100	100	100	100
Ammaniti	100	100	100	100	100	100	100	100	100	100
Veronesi	75	75	100	100	100	75	75.5	100	100	100
Lagioia	100	100	100	100	100	100	100	100	100	100
Bajani	100	100	100	100	100	100	100	100	100	100
Rea	100	100	100	100	100	100	100	100	100	100
Benni	100	100	100	100	100	100	100	100	100	100
Pincio	67	100	100	100	100	67	100	100	100	100
Ramondino	100	100	100	100	100	100	100	100	100	100
Brizzi	67	100	100	100	100	67	100	100	100	100
Montesano	100	100	100	100	100	100	100	100	100	100
Prisco	100	100	100	100	100	100	100	100	100	100
Balzano	100	100	100	100	100	100	100	100	100	100
Affinati	100	100	100	100	100	100	100	100	100	100
Milone	100	100	100	100	100	100	100	100	100	100
Vasta	100	100	100	100	100	100	100	100	100	100
Covacich	100	100	100	100	100	100	100	100	100	100
Morazzoni	50	100	50	100	100	50	100	50	100	100
Maraini	75	75	88	88	88	93	93	93	93	93

Baricco	50	100	100	100	88	50	100	100	100	100
Sereni	83	83	83	83	83	83	83	83	83	83
Nori	33	33	56	67	67	33	0	56	67	67
Fois	67	67	67	67	67	67	67	67	67	67
Nesi	0	0	33	0	33	0	0	33	0	67
Valerio	33	33	67	67	67	33	33	50	50	50
Vinci	50	50	50	50	50	50	50	50	50	50
Raimo	0	50	50	50	50	0	50	50	50	50
Murgia	40	0	40	0	40	40	0	40	0	40
Giordano	0	33	33	67	33	0	33	33	67	33
Parrella	0	0	0	0	0	0	0	0	0	0
Scarpa	0	0	0	0	0	0	0	0	0	0

More explicitly, we can ask: if the method incorrectly attributes more than one text of a given author, the different texts are all attributed to the same (incorrect) author or the attribution is spread among different authors? The latter case is an indication of a lack of a clear stylistic signature, or a great eclecticism, of the unknown author. Secondly, when using the majority rules on fragments, how the attribution of the different fragments is distributed? To quantify the answers to these questions we shall compute, for each author, say Y, in the corpus, the entropy of the distribution of the authors associated to Y by our algorithm. We will do that both when attributing the entire texts, and when attributing single fragments. More explicitly, let us focus on the texts (respectively fragments) of the author Y, and let us call t_Z (respectively f_Z for fragments) the number of times an author Z of the corpus is chosen by the algorithm to be the author of a text (a fragment) of Y. By computing, based on the number of occurrences t_Z (respectively f_Z for fragments), the probabilities p_Z that a generic author Z is chosen as the author of the texts (fragments) of Y, we can write the Shannon entropy (Shannon, 1948) for the distribution of the authors associated by our method to Y:

$$H_Y = -\sum_{i=1}^{N} p_i \log p_i$$

where here N is the number of different authors in the corpus (in our case N=40) and represents the probability that works of the author Y are attributed to the generic author *i*. It is well known that $H_Y = 0$ if and only if all the probabilities, but one, are zero, i.e., a situation of certainty (or maximum order), in our case corresponding to the attribution of all the texts (fragments) of Y to a single author (not necessarily the correct one). Otherwise, H_Y is positive and achieve its maximum value $H_{max} = \log(N)$ when events are all equiprobable, i.e., the most uncertain situation (or maximum disorder). We note that if we had a sufficiently

high number of texts (or fragments), the maximum observed value of the entropy would be log(*N*). In most of our cases, however, the number of texts or of fragments are smaller than the number of the authors in the corpus: in those cases the maximum possible value for the entropy is log(T_Y) (respectively log(F_Y)), where we call T_Y and F_Y respectively the number of texts and of fragments of the author Y. To fix the idea, let us compute explicitly the entropy for De Silva (author Y), both for texts (T_Y = 8) and for fragments (F_Y = 80). Since all De Silva's texts are correctly attributed, the entropy computed for texts, is zero. On the other hand, for the fragments, 77 of them are attributed to De Silva himself and 3 to Lagioia (in this case Lagioia would be one of the authors Z). The entropy thus reads:

$$p_1 = 0.9625 \qquad p_2 = 0.0375 \qquad p_3 = \cdots = p_{40} = 0$$

$$H^{(DS)} = -p_1 \log(p_1) - p_2 \log(p_2) = 0.2307$$

In order to compare the values of the entropy for different authors, with different number of texts and fragments, we define the normalized entropy, that is the entropy divided by its maximum value, that in the example above reads:

$$h^{(DS)} = \frac{H^{(DS)}}{\mathrm{H_{max}}} = \frac{H^{(DS)}}{\log(40)} = 0.0434$$

All entropy evaluations are carried out for extended corpus, since this analysis lies outside the attribution aim and, at the same time, takes deep advantage of a more accurate statistics.

As discussed for the robustness analysis, in evaluating the Starnone's entropy we excluded from the corpus all the texts of Elena Ferrante, and viceversa, when evaluating Ferrante's entropy, we excluded from the corpus all the texts of Starnone.

In Figure 3, entropy values for each author, both relative to the attribution of text and fragments, are shown together with success rates obtained in a WMR_90 scheme of each author, which we recall to be corresponding to a Weighted-Profile Majority Rule analysis conducted on first 90 fragments in the ranking of the test fragment. It is worth noting that, although is value is not independent from the success rate, it still gives an additional information. In order to clarify this point, we survey Parrella's case. A WMR_90 scheme does not allow to attribute correctly any of his texts, and we could expect as a consequence a high text-entropy value. However, the latter is zero since both of Parrella's texts are ascribed to the writer Balzano. This result could mark a specific similarity between these authors.

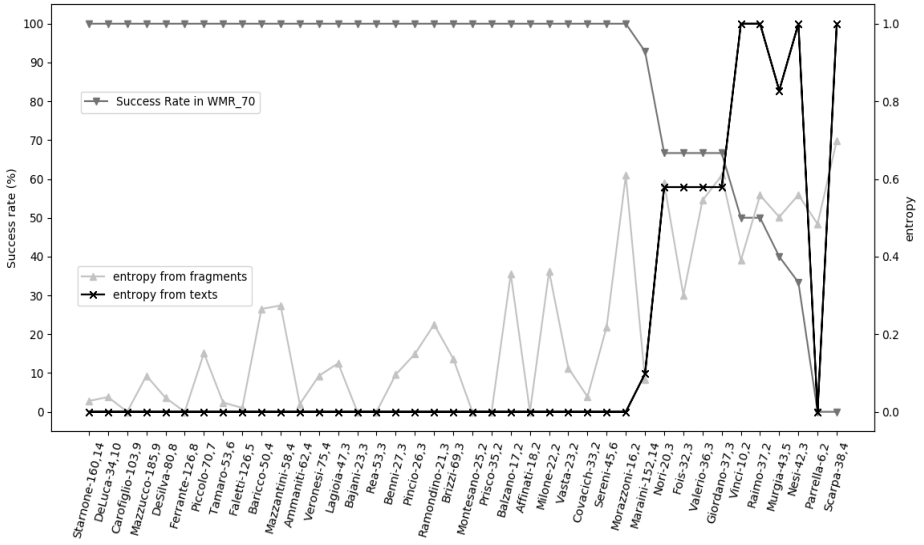

Figure 3. Entropy values for each author, together with his success rate in a WMR_90 scheme, are shown. X ticks outline every writer and, respectively, the number of fragments and texts participating the extended corpus.

Attribution

This section is devoted to the specific problem of identifying the most likely name behind Elena Ferrante's work. Here we then compare each text signed by Ferrante with all the other in the reference corpus (both the original and the extended one).

In Table 4 and Table 5 we show results of the attribution using both the Majority Rule scheme and the Weighted Majority Rule scheme for fragments, that are the outperforming algorithms previously introduced. The outcome seems to be extremely clear: between Ferrante's texts, eight out of eight are attributed to Domenico Starnone.

A captivating task is to perform the same attribution including Ferrante's texts themselves in Ferrante's ranking. So that an attribution in Starnone's direction would imply an attribution to Starnone *rather than* to Ferrante. Results are shown in Table 6 and Table 7.

As a note if interest, we performed the attribution in the opposite direction, aiming to identify the author behind texts signed by Starnone. At odds with the results presented in the previous section, we now include the texts of Ferrante in the reference corpus, and we investigate both the cases in which texts of Starnone himself are included and excluded from the corpus. In the former case 3 texts out of 14 are attributed to Ferrante, in the latter case 11 (see Table 8).

Table 4. Outcome of attribution within original corpus (Ferrante's text excluded) explicitly for each text signed by Elena Ferrante.

	Authorship attribution of Elena Ferrante (original corpus)					
	29566 bytes fragments		20000 bytes fragments		10000 bytes fragments	
	MR	**WMR_90**	**MR**	**WMR_90**	**MR**	**WMR_90**
L'amore molesto	Starnone	Starnone	Starnone	Starnone	Starnone	Starnone
I giorni dell'abbandono	Starnone	Starnone	Starnone	Starnone	Starnone	Starnone
La figlia oscura	Starnone	Starnone	Starnone	Starnone	Starnone	Starnone
L'amica geniale	Starnone	Starnone	Starnone	Starnone	Starnone	Starnone
Storia del nuovo cognome	Starnone	Starnone	Starnone	Starnone	Starnone	Starnone
Storia di chi fugge e di chi resta	Starnone	Starnone	Starnone	Starnone	Starnone	Starnone
Storia della bambina perduta	Starnone	Starnone	Starnone	Starnone	Starnone	Starnone

Table 5. Outcome of attribution within extended corpus (Ferrante excluded), explicitly for each text signed by Elena Ferrante.

	Authorship attribution of Elena Ferrante (extended corpus)					
	29566 bytes fragments		20000 bytes fragments		10000 bytes fragments	
	MR	**WMR_90**	**MR**	**WMR_90**	**MR**	**WMR_90**
L'amore molesto	Starnone	Starnone	Starnone	Starnone	Starnone	Starnone
I giorni dell'abbandono	Starnone	Starnone	Starnone	Starnone	Starnone	Starnone
La Frantumaglia	Starnone	Starnone	Starnone	Starnone	Starnone Carofiglio	Starnone Carofiglio
La figlia oscura	Starnone	Starnone	Starnone	Starnone	Starnone	Starnone
L'amica geniale	Starnone	Starnone	Starnone	Starnone	Starnone	Starnone
Storia del nuovo cognome	Starnone	Starnone	Starnone	Starnone	Starnone	Starnone
Storia di chi fugge e di chi resta	Starnone	Starnone	Starnone	Starnone	Starnone	Starnone
Storia della bambina perduta	Starnone	Starnone	Starnone	Starnone	Starnone	Starnone

Table 6. Outcome of attribution within original corpus (Ferrante's texts included), explicitly for each text signed by Elena Ferrante. We highlight in bold the occurrences of the name Starnone, i.e., cases where a work from Ferrante is attributed to Starnone instead to Ferrante himself/herself.

	Authorship attribution of Elena Ferrante (original corpus)					
	29566 bytes fragments		20000 bytes fragments		10000 bytes fragments	
	MR	**WMR_90**	**MR**	**WMR_90**	**MR**	**WMR_90**
L'amore molesto	Starnone	Starnone	Starnone	Starnone	Starnone	Starnone
I giorni dell'abbandono	Ferrante	Ferrante	Ferrante	Ferrante	Ferrante	Ferrante
La figlia oscura	Ferrante	Ferrante	Ferrante	Ferrante	Ferrante	Ferrante
L'amica geniale	Ferrante	Ferrante	Ferrante	Ferrante	Ferrante	Ferrante
Storia del nuovo cognome	Ferrante	Ferrante	Ferrante	Ferrante	Ferrante	Ferrante
Storia di chi fugge e di chi resta	Ferrante	Ferrante	Ferrante	Ferrante	Ferrante	Ferrante
Storia della bambina perduta	Ferrante	Ferrante	Ferrante	Ferrante	Ferrante	Ferrante

Table 7. Outcome of attribution within extended corpus (Ferrante's texts included), explicitly for each text signed by Elena Ferrante. We highlight in bold the occurrences of the name Starnone, i.e., cases where a work from Ferrante is attributed to Starnone instead to Ferrante himself/herself.

	Authorship attribution of Elena Ferrante (extended corpus)					
	29566 bytes fragments		20000 bytes fragments		10000 bytes fragments	
	MR	**WMR_90**	**MR**	**WMR_90**	**MR**	**WMR_90**
L'amore molesto	Starnone	Starnone	Starnone	Starnone	Starnone	Starnone
I giorni dell'abbandono	Ferrante	Ferrante	Ferrante	Ferrante	Ferrante	Ferrante
La Frantumaglia	Ferrante	Ferrante	Starnone	Ferrante	Ferrante	Ferrante Valerio
La figlia oscura	Ferrante	Ferrante	Ferrante	Ferrante	Ferrante	Ferrante
L'amica geniale	Ferrante	Ferrante	Ferrante	Ferrante	Ferrante	Ferrante
Storia del nuovo cognome	Ferrante	Ferrante	Ferrante	Ferrante	Ferrante	Ferrante

Storia di chi fugge e di chi resta	Ferrante	Ferrante	Ferrante	Ferrante	Ferrante	Ferrante
Storia della bambina perduta	Ferrante	Ferrante	Ferrante	Ferrante	Ferrante	Ferrante

Table 8. Outcome of the attribution of texts signed by Starnone in two cases: including Starnone between candidate authors and excluding him. The attribution is performed for fragments of texts 29566 bytes long and in a Weighted Majority Rule scheme for first 90 ranked fragments. We highlight in bold the occurrences of the name Ferrante, i.e., cases where a work from Starnone is attributed to Ferrante instead to Starnone himself.

	Starnone excluded	Starnone included
Ex cattedra	Affinati	Starnone
Il salto con le aste	Raimo	Starnone
Fuori registro	**Ferrante**	Starnone
Eccesso di zelo	**Ferrante**	**Ferrante**
Denti	**Ferrante**	Starnone
Via Gemito	**Ferrante**	**Ferrante**
Labilità	**Ferrante**	Starnone
Prima esecuzione	**Ferrante**	**Ferrante**
Condom Butterfly	Raimo	Starnone
Fare scene. Storie di cinema	**Ferrante**	Starnone
Spavento	**Ferrante**	Starnone
Autobiografia erotica di Aristide Gambia	**Ferrante**	Starnone
Lacci	**Ferrante**	Starnone
Scherzetto	**Ferrante**	Starnone

For the sake of completeness, in Table 9 we show the normalized entropy values for Ferrante's and Starnone's fragments in two cases: Ferrante and Starnone neglected, respectively, in Starnone's and Ferrante's ranking (namely the conditions imposed in section 4.1); Ferrante and Starnone included in each other's ranking. In the first case the author entropy quantifies self-detectability, or auto-similarity. By comparing this value with the one obtained in the latter case, we gain one more tip on the strong connection between these two authors, or, more properly, on a connection much stronger than the one between all other authors in equipped corpus.

Table 9. Normalized entropies of Ferrante's and Starnone's fragments in two cases. Starting from the left: O-E-E: the entropy of each one is obtained excluding the fragments of the other writer from the corpus; O-I-E: the entropy of each one is obtained including the fragments of the other writer in the corpus.

	Other-Excluding-Entropy	Other-Including-Entropy	Number of Fragments
Ferrante	0.000	0.076	126
Starnone	0.039	0.190	160

Conclusions

In this final section we summarize the content of the paper and its main results. We presented a data-compression oriented technique (Lempel and Ziv, 1977; Grumbach and Tahi, 1994; Loewenstern *et al.*, 1995; Li *et al.*, 2001; Benedetto *et al.*, 2002; Baronchelli *et al.*, 2005) through which it is possible to quantify the similarity between two generic sequences of characters, in particular texts. The method allowed to solve for instance the *Grunberg–Van der Jagt* problem in The Netherlands[4]. We tested the method to the corpus of 150 texts considered in this book as well as to an extended version of it. The results for authorship attribution are very good featuring an overall rate of success slightly below 90%. We then considered the attribution of the works of Elena Ferrante. Some of us already considered the matter in 2006[5], well before the appearance of Ferrante's tetralogy of *L'amica geniale,* to conclude for a strong similarity between the work of Ferrante and that of Domenico Starnone. The same conclusions have been reached in the more recent analyses performed with the corpus considered in this book. All these analyses point to the same conclusion of a very strong similarity between Starnone and Ferrante, so strong that often works of Ferrante are erroneously attributed to Starnone (instead of Ferrante) and viceversa. From this similarity the only reasonable conclusion to be drawn is that, within the corpus considered (both the original and the extended one) the most likely author behind the *nom de plume* of Elena Ferrante is Domenico Starnone. We remark again that, despite the strong hints pointing to Starnone, we cannot rule out the possibility that Ferrante's signature could hide another author (or several authors) not included in the specific corpus considered. There has been a strong interest in that matter on the international newspapers. Though this

[4] Grunberg–Van der Jagt authorship attribution problem: https://www.nrc.nl/nieus/2002/05/11/grunberg-is-van-der-jagt-7589390-a1067475

[5] L. Galella, *Ferrante è Starnone. Parola di computer*, L'Unità, 23 November 2006.

is not the place to report all the different hypotheses made, it is interesting to mention the hypothesis that Elena Ferrante is actually hiding the translator and author Anita Raja[6]. Here we can only say that, since Anita Raja is not a novelist, her production being composed by translations and essays, she could not be included in the same corpus analyzed here. More specific analyses are still ongoing in order to shed light on this last point.

References

Badii, R. and Politi, A. (1997). *Complexity, Hierarchical Structures and Scaling in Physics*. Cambridge: Cambridge University Press.

Baronchelli, A., Caglioti, E. and Loreto, V. (2005). Artificial sequences and complexity measures, *Journal of Statistical Mechanics: Theory and Experiment*, 4, P04002.

Bell, T.C., Cleary, J.C. and Witten, I.H. (1990). *Text Compression*. Englewood Cliffs: Prentice-Hall.

Benci, V., Bonanno, C., Galatolo, S., Menconi, G. and Virgilio, M. (2002). *Dynamical systems and computable information, Discrete and Continuous Dynamical Systems* B 4 935 and references therein [cond-mat/0210654].

Benedetto, D., Caglioti, E., and Loreto, V. (2002). Language trees and zipping, *Physical Review Letters* 88, 048702-2 – 048702-4.

Boffetta, G., Cencini, M., Falcioni, M. and Vulpiani, A. (2002). Predictability: a way to characterize complexity, *Physics Reports*, 356, 367-476.

Chaitin, G.J. (1966). On the length of programs for computing finite binary sequences, *Journal of the Association for Computing Machinery* 13, 547-569.

Chaitin, G.J. (1990). *Information, Randomness and Incompleteness*. Singapore: World Scientific.

Eder, M. (2017). Short samples in authorship attribution: A new approach. In *Digital Humanities 2017, Conference Abstract*. Montreal: McGill University, 221–24.

El-Yaniv, R., Fine, S. and Tishby, N. (1997). Agnostic classification of markovian sequences, *Advances in Neural Information Processing Systems*, 10, 465-471.

Farach, M., Noordewier, M., Savari, S., Shepp, L., Wyner, A. and Ziv, J. (1995). On the entropy of DNA: algorithms and measurements based on memory and rapid convergence. In *Proceedings of the sixth annual ACM-SIAM symposium on Discrete algorithms* (SODA '95). Philadelphia: Society for Industrial and Applied Mathematics, 48-57.

[6] http://www.nybooks.com/daily/2016/10/02/elena-ferrante-an-answer/

Grumbach, S. and Tahi, F. (1994). A new challenge for compression algorithms: genetic sequences. *Information Processing & Management*, 30, 875-886.

Juola, P. (1998). Cross-entropy and linguistic typology. In Powers, D.M.W. (ed.), *NeMLaP3/CoNLL98: New Methods in Language Processing and Computational Natural Language Learning*. Sydney: ACL, 141-149.

Juola, P. and Baayen, H. (2005). A Controlled-Corpus Experiment in Authorship Attribution by Cross-Entropy, *Literary and Linguistic Computing*, 20(1), 59-67.

Khinchin, A.I. (1957). *Mathematical Foundations of Information Theory*. New York: Dover Publications.

Kolmogorov, A.N. (1965). Three approaches to the quantitative definition of information, *Problems of Information Transmission*, 1, 1-7.

Kukushkina, O.V., Polikarpov, A.A. and Khmelev, D.V. (2000). Using Literal and Grammatical Statistics for Authorship Attribution, *Problemy Peredachi Informatsii*, 37, 96-108 (in Russian; translated in English, in *Problems of Information Transmission*, 37 (2001), 172-184).

Lempel, A. and Ziv, J. (1977). A Universal Algorithm for Sequential Data Compression, *IEEE Transactions on Information Theory*, 23(3), 337-343.

Li, M., Badger, J., Chen, X., Kwong, S., Kearney, P. and Zhang, H. (2001). An information-based sequence distance and its application to whole mitochondrial genome phylogeny, *Bioinformatics*, 17(2), 149-154.

Li, M. and Vitányi, P. (1997). *An Introduction to Kolmogorov Complexity and Its Applications*. New York: Springer.

Lind, D. and Marcus, B. (1995). *Symbolic Dynamic and Coding*. Cambridge: Cambridge University Press.

Loewenstern, D., Hirsh, H., Yianilos, P. and Noordewieret, M. (1995). DNA sequence classification using compression-basedinduction, *DIMACS Technical Report*, 95-104.

Milosavljević, A. (1995). Discovering Dependencies via Algorithmic Mutual Information: A Case Study in DNA Sequence Comparisons, *Machine Learning*, 21(1-2), 35-50.

Puglisi, A., Benedetto, D., Caglioti, E. and Loreto, V. (2003). Data-compression and learning in time sequences analysis, *Physica D: Nonlinear Phenomena* 180 (1-2), 92-107.

Shannon, C. E. (1948). A Mathematical Theory of Communication: Part II, The Discrete Channel with Noise, *The Bell System Technical Journal*, 27, 623-656.

Solomonoff, R.J. (1964). A formal theory of inductive inference, *Information and Control*, 7, 1-22, 224-254.

Stamatatos, E. (2009). A Survey of Modern Authorship Attribution Methods, *Journal of the American Society for Information Science and Technology*, 60(3), 538-556.

Teahan, W.J. (2000). Proceedings of the International Conference on Content-based Multimedia Information Access (RIAO 2000), CID-CASIS, Paris, 943-961.

Thaper, N. (2001). *MS in Computer Science*. Master's Thesis. Cambridge: MIT Press.

Verdú, S. (1998). Fifty years of Shannon theory, *IEEE Transactions on Information Theory*, 44(6), 2057-2078.

Wyner, A.D. and Ziv, J. (1994). The sliding-window Lempel-Ziv algorithm is asymptotically optimal, *Proceedings of the IEEE*, 82(6), 872-877.

Ziv, J. and Merhav, N. (1993). A measure of relative entropy between individual sequences with application to universal classification, *IEEE Transactions on Information Theory*, 39, 1270.

Zurek, W. H. (1990, ed.). *Complexity, Entropy, and Physics of Information*. Redwood City, Addison-Wesley.

Blended Authorship Attribution: Unmasking Elena Ferrante Combining Different Author Profiling Methods

George K. Mikros

National and Kapodistrian University of Athens, Greece

Abstract

The aim of this paper is to explore the authorship of Elena Ferrante's novels using a cascade of author profiling methods applied to different author's characteristics (gender, age, town and region he/she grew). The method proposed combines the above-mentioned profiling tasks with standard authorship attribution methodology and can be considered as a novel approach to authorship verification problems. All experiments have been conducted using a rich document representation schema, the Author's Multilevel Ngram Profiles (AMNP) consisting of character and word ngrams of increasing length (n= 1-3). AMNP was used to train a robust machine learning algorithm (SVM with polynomial kernel) and all profiling results were highly accurate (over 90%) indicating that the person behind Ferrante is a male, aged over 60, from the region Campania and the town Saviano. The combination of these characteristics indicate a single candidate (among the authors of our corpus), Domenico Starnone.

Introduction

Authorship identification refers to the connection of a text of unknown authorship to a specific author using a set of quantifiable text features as indicators of the author's style. Modern stylometric methods are based increasingly in advanced machine learning methods and a variety of textual features in order to identify each author's style. Major landmarks in the field were the authorship

analysis of *The Federalist Papers* performed by Mosteller and Wallace (1984) and the multivariate statistical methods introduced by Burrows (Burrows, 1987, 1989, 1992) and his associates (Burrows and Craig, 1994; Burrows and Hassal, 1988).

Since the late 1990s authorship identification has known a new impetus based on developments in a number of key research areas such as Information Retrieval, Machine Learning and Natural Language Processing (Stamatatos, 2009). Furthermore, online text is now massively available, and Web 2.0 has added to the now standard internet genres of email, web page and online forum message, new forms of online expression such as blogs, tweets and instant messaging.

Language usage was long been recognized as a carrier of various extralinguistic information such as historical period, dialect (both geographical and social), author's gender and age, ideology etc. Using more or less the same experimental setup, we can identify not only the author's identity, but also various author's characteristics both biological and psychological. This kind of analysis, called author profiling, is gaining interest in the research community (see Reddy, Vardhan, and Reddy, 2016 for a current literature review) as its possible applications are wider than the standard authorship attribution. Moreover, profiling characteristics can be combined, and we can relate a text with its author across multiple dimensions that can uncover deep links between linguistic production and aspects of our biological and psychological being.

Authorship identification avoiding false positive results

An interesting application of authorship identification methodology appears in the case of Elena Ferrante. Elena Ferrante is a well-known Italian novelist with international fame. Her/his books are best sellers in many countries and their literary value is now widely accepted. Since her/his first book publication in 1992, her/his identity has been kept secret and until now, remains an open question. She has published 7 novels than have been translated in many languages. She represents a real-life challenge for computational stylistic methods and related techniques.

However, in the literature of authorship identification, most of the published research has been directed with datasets that are constructed for the specific experiments. The real-life stylometric attributions are not so many since the discipline has been suffered by various misattributions and examples of malpractice, e.g. the CUSUM controversy (Canter, 1992; Hardcastle, 1997; Sanford *et al.*, 1994) or the misattribution of the sonnet "Elegy" to Shakespeare by Donald Foster (Foster, 1989). Since stylometric methods don't have a standard error rate yet, real-life attributions should obey what Smith (1990, pp. 249-250) has described as the six principles of literary attribution:

1. The onus of the proof lies entirely with the person making the ascription.
2. The argument of adding something to an author's canon has to be vastly more stringent than for keeping it there.
3. If doubt persists, an anonymous work must remain anonymous.
4. Avoidance of a false attribution is far more important than failing to recognize a correct one.
5. Only works of known authorship are suitable as a basis for attributing a disputed work.
6. There are no short-cuts in attribution studies.

The above six points define a very strict framework for conducting real life authorship attribution studies and they should always be considered when we are dealing with authors whose works have impact to millions of readers. In a typical authorship attribution study, we would address directly this problem by training our algorithms in a closed set of candidate authors and letting them decide whether the anonymous texts belong at least to one of the included possible authors. The algorithm would be forced to come up with an author's name even when the texts were not written by any of those, maximizing type I errors in the experimental design, e.g. conflicting directly the Smith's points 2, 3, and 4.

However, in this study we are planning to tackle the problem from another angle. We will train separate author profiling models using the metadata available in the training corpus of modern Italian authors, that is, gender, age, region and city. Each model will identify Ferrante's identity in terms of gender, age, region and city. These information combined can narrow down the sample space of candidate authors and restrict (or match) the candidate author(s) that share these characteristics.

Our experimental analysis has been developed so that it will respect Smith's principles. Instead of using very specific categories (authors' names), we will use the broader author's characteristics that can profile wider, open sets of candidates and can help us identify the author's main identity dimensions. This approach can help us approach open-class authorship problems since it is not restricted to the authors' characteristics available in the training sample but can go beyond that. An author's identity that can be defined across 2-3 major features (such as gender, age and region) can be used to link an author to a specific name among the candidate authors inside the training corpus, but at the same time can be used as a generic profile where different analysts could also make alternative hypotheses proposing possible authors that are not inside the training data. We are not forcing a name out of this procedure, but we describe a profile and the match can or can't occur.

Materials and methods

Corpus

This study is based on a literary corpus of modern Italian novels carefully prepared by Michele Cortelazzo and Arjuna Tuzzi from the University of Padua (Tuzzi and Cortelazzo, 2018). The corpus contains 150 novels from 40 different authors, most of them written between 1987 and 2016 totaling 9,837,851 tokens. The corpus consists of texts of variable length (Min= 8,129 words, Max= 194,993 words, St.Dev= 38,366) and Ferrante is represented by all her 7 novels (623,466 tokens) which are also variable in size (Min= 36,091 words, Max= 139,491 words, St.Dev.=45,282). It contains not only the authors that have been suspected to be behind Ferrante's name, but also a wider range of authors that give a wider picture of the literary production of modern Italian literature. In that sense, the specific corpus can be used to explore Ferrante's position in the wider framework of modern Italian literature and can be used to model author profiles that have a more generic coverage.

All books were converted in plain text files with UTF-8 encoding. In order to increase our sample space and enhance our machine learning modeling we sliced each novel in 1,000 words chunks increasing our vector sample from 150 to 9,514 cases.

Stylometric Features

In order to train effectively our profile models, we developed a feature-rich document representation model comprised by the following features groups:

1. Author Multilevel N-gram Profiles (AMNP): 3,000 features, 1,000 features of each n-gram category (2-grams and 3-grams at the character level, and 2-grams at the word level);
2. Most Frequent Words in the corpus (MFW): 1,000 features.

The first feature group (AMNP) provides a robust document representation which is language independent and able to capture various aspects of stylistic textual information. It has been used with success in authorship attribution problems (Mikros and Perifanos, 2011; 2013) and gender identification focused on bigger texts (e.g. blog posts, see Mikros, 2013). AMNP consists of increasing order n-grams in both character and word level. Since character and word n-grams capture different linguistic entities and function complementary, we constructed a combined profile of 2, 3 characters n-grams and 2 words n-grams. For each n-gram we calculated its normalized frequency in the corpus and included the 1,000 most frequent entries resulting in a combined vector of 3,000 features. AMNP due to its linguistic multilevel character, it's extremely flexible document representation. Coupled with SVM can adjust in different classifica-

tion tasks since the SVM each time will make a different feature selection in order to extract the support vectors of each classification. Thus, each time SVM will use different subpart of AMNP and each time this subpart will be the optimum for each classification.

The second feature group (MFW) can be considered classic in the stylometric tradition and it is based on the idea that the MFWs belong to the functional words class and are beyond the conscious control of the author, thus revealing its stylometric fingerprint. In this study we used the 1,000 most frequent words of the corpus.

Machine Learning Algorithms

The above described features have been exploited for training a classification machine learning algorithm, Support Vector Machines (SVM) (Vapnik, 1995), in four different author profiling tasks (author's gender, age, and geographical region and town). SVM is considered a state-of-the-art algorithm for text classification tasks (Diederich *et al.*, 2003; Joachims, 1998). The SVM constructs hyper-planes of the feature space in order to provide a linear solution to the classification problem. For our trials we experimented with various kernels and we ended up choosing the polynomial one as this was the most accurate in our dataset. All statistical models developed have been evaluated using 10-fold cross validation (90% training set – 10% testing set) and the accuracies reported represent the mean of the accuracies obtained in each fold. Since the feature space was sparse, we eliminated all features that showed a variance close to zero, using the two following rules: the percentage of unique values was less than 20%, and the ratio of the most frequent to the second most frequent value was greater than 20. The near-zero variance feature removal shrank the number of the employed features and led to a reduction of 33.2% (from the initial 4,000 available features we kept 2,672 features). Moreover, since SVM optimization occurs by minimizing the decision vector w, the optimal hyperplane is influenced by the scale of the input features and for this reason we standardized the data (z-scores with mean 0 and variance 1) prior to SVM model training.

Results

SVM models with polynomial kernel have three hyper-parameters that influence the learning process of the algorithm and its generalization power, i.e. degree, scale and C-parameter. These hyper-parameters have to be empirically adjusted since they are dependent on the specific training dataset. In order to train the optimum models for the profiling tasks we used hyper-parameter

tuning exploiting the automatic grid method provided by the *Caret* R package (Kuhn *et al.*, 2012) and setting 3 different values per hyper-parameter. This tuning process created 3^3= 27 models (one for each combination of the 3 hyper-parameter values) and we kept the model with the best prediction cross-validated accuracy.

In the Gender profiling model, the best model obtained cross-validated accuracy of 93.6%. The confusion matrix of this classification can be found in Table 1:

Table 1. Confusion matrix for the gender profiling task

Reference Prediction	Female	Male
Female	**703**	54
Male	134	**2074**

The model performs rather asymmetrically across the two genders since it predicts females with less recall (0.84) than males (0.975) but the precision for females and males is comparable (0.929 and 0.93 correspondingly). Its overall accuracy is high (0.936). We used this model to predict the gender of each of the Ferrante's text chunks. The result was that 594 of the 619 chunks (96%) belong to a male author. Since our model is extremely sensitive to male authorship (exhibits high precision and recall to this category), the gender profiling should be considered highly reliable.

The second profiling model we developed was about the age of the author. The provided corpus contained the age of each author. However, since we are training classification models, we had to transform the numerical variable to qualitative. More specifically, we merged the ages in 3 age-groups, i.e. less than 40 years old represented with 4 authors, between 40 and 60 years old represented with 24 authors and older than 60 represented with 11 authors. The developed model obtained a cross-validated accuracy of 0.93 and its confusion matrix can be found below (Table 2):

Table 2. Confusion matrix for the age profiling task

Reference Prediction	< 40 years	40 – 60 years	> 60 years
< 40 years	**38**	0	1
40 – 60 years	15	**540**	27
> 60 years	3	14	**252**

The confusion matrix reveals that the model is very accurate overall. However, its accuracy is different across the age groups. More specifically, the age category < 40 exhibits low recall ~ high precision (recall= 0.679, prec=97.4) meaning

that it is very picky. When the model identify a < 40 age category author, then it usually is right, but at the same time it misses a lot of other texts written by authors of this age category. However, in all the other age categories the model exhibits high precision ~ high recall (40 – 60 years, recall= 0.975, prec= 0.93, > 60 years, recall= 0.9, prec= 0.937). This model predicted that Ferrante is over 60 in 561 out of 619 chunks (91%), an age category that is identified with high recall and precision in our training data.

A third profiling model was developed for the author's region. The available authors were derived from 11 different Italian regions. Campania and Lazio are the home areas of half of the authors (10 and 9 authors respectively). The best model obtained yielded 0.9 cross-validated accuracy and its confusion matrix can be found below (Table 3):

Table 3. Confusion matrix for the region profiling task

Reference Prediction	Campania	Emilia Romagna	Friuli V.G.	Lazio	Lombardia	Piemonte	Puglia	Sardegna	Sicilia	Toscana	Veneto
Campania	**665**	11	4	30	11	14	7	8	1	15	8
Emilia Romagna	0	**165**	1	1	0	1	0	0	0	1	0
Friuli V.G.	0	0	**114**	0	0	1	0	0	0	0	1
Lazio	37	13	12	**671**	11	16	10	8	3	23	15
Lombardia	0	0	0	1	**51**	0	0	0	0	0	0
Piemonte	4	0	1	2	1	**318**	1	1	0	0	0
Puglia	2	0	0	2	0	0	**227**	1	0	1	0
Sardegna	0	0	0	1	0	0	0	**103**	0	0	0
Sicilia	0	0	0	0	0	0	0	0	**35**	0	0
Toscana	2	4	1	4	0	2	0	3	0	**283**	1
Veneto	0	0	0	0	0	0	0	0	0	0	**35**

Analyzing the confusion matrix we can see that there is significant variation among the prediction accuracies in various regions. Veneto and Sicilia have 1, while the lowest accuracy is observed in Lazio (0.82) and Campania (0.86). It's interesting that these two areas are neighboring and exhibit the biggest number of misclassifications between them. A possible explanation would be that Lazio contains the Italy's capital Rome, where many authors from different places have stayed and worked. The idiolects contained in this region label are heterogeneous and could be mixed idiolects from various other places. In fact the confusion matrix gives a high dispersion of misclassification errors across all region labels and Lazio. Moreover, both Campania and Lazio contain the biggest

number of text chunks increasing the probability of errors in the classification task. The Region model predicted that Ferrante's region is Campania in 607 out of 619 chunks (98%).

A last profiling model was developed using the author's born town. Although, this classification label correlates with the region (town labels are 22 in total and are sub-part of the region labels), it is more detailed and splits the authors in more equal groups. The best model achieved cross-validated accuracy of 0.93 which is even better that the accuracy obtained for the region classification. When applied to Ferrante's data predicts that the author's town is Saviano in 608 from 619 chunks (98%).

Combining the predictions of all four profiling models we created a matrix of author characteristics that filters the initial pool of available authors and step by step drives us to the most probable candidate. A visualization of the path is depicted in Figure 1 below:

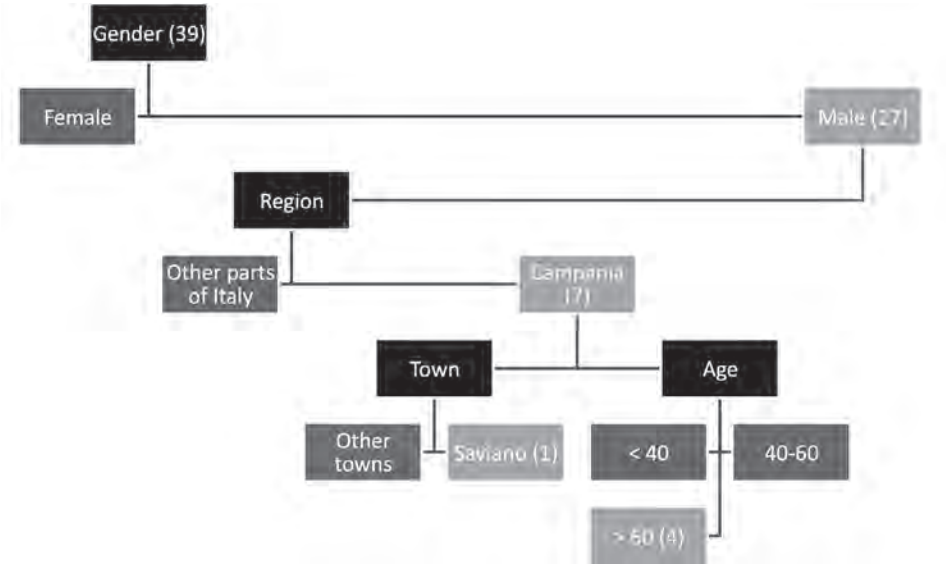

Figure 1. Navigating the profiling restrictions. In parentheses the numbers of suspected authors

The first filter applied is the author's gender. Our model predicted that Ferrante is a male author and this restricted the initial sample of 40 authors to 27 males. The second filter we apply is the Region. Our model predicted that Ferrante is from Campania and this restricts further the 27 male authors to 7. If we apply sequentially the predictions of the town model (Saviano), the only author in our corpus that satisfies all these three characteristics is Domenico Starnone.

Alternatively, instead of using the town model, we could use the age profiling model which has predicted that Ferrante is over 60. Applying this, restricts the Ferrante's candidate authors to 4 (De Luca, Prisco, Rea, Starnone). In that case, since we don't have other profiling model to combine, we can just run a standard authorship attribution study and consider these four candidate authors as a closed group. In fact, we used the same experimental setup (4,000 features – AMNP and MFW and the SVM with polynomial kernel) using the 4 candidate authors and the cross-validated accuracy of the model was perfect (1), i.e. the model could predict perfectly if any text chunk of these 4 authors had been written by whom. This model then was applied to Ferrante's data and it predicted that all Ferrante's chunks (619 of 619) have been written by Domenico Starnone.

Conclusion

In this study we presented a blended authorship attribution method where multiple author profiling classifications restricted the initial sample of candidate authors to a few or even the one most probable real author. This method is generic in the sense that we developed a frame of author's characteristics that can reliably identify the real author even when he/she is not among the candidates.

We developed two different scenarios working with this method. The first was based on cascading profiles that were combined so that a single candidate emerged as the real author behind Elena Ferrante. More specifically, we combined the gender, the Region and the town profiles and Ferrante was identified as a male author, from Campania and more specifically from Saviano. These characteristics give as a single candidate among the 39 candidate authors in our corpus, Domenico Starnone. The second scenario was based on using the multiple profiling models as a filter in our initial pool of candidate authors. In this case, the initial sample of authors will be significantly restricted in a handful closed set of candidate authors. In this set we can apply reliably the standard authorship attribution methodology, since the pool of candidate authors will be small and we have increased confidence (due to the profiling restrictions) that it is closed (i.e. we are certain that at least one of them is the real author). In both scenarios, Domenico Starnone emerged as the most probable author behind the pseudonym Elena Ferrante.

The proposed method can be helpful in real-life semi-open authorship identification problems. It respects Smith's criteria about increased protection against Type I errors in our experimental methodologies. Moreover, it helps us transform an open authorship verification problem to a closed authorship attri-

bution one since it can limit an unrestricted, open sample-space of candidates to a small, well-defined closed group. Authors' characteristics such as the gender and the age are generic and can be used for this purpose effectively.

Acknowledgments

The author would like to express his gratitude to Arjuna Tuzzi and Michele Cortelazzo since they inspired and supported this study in many ways and to the University of Padova that funded author's stay in Padua during the IQ-LA-GIAT Summer School in Quantitative Analysis of Textual Data (3rd Edition – September 2017).

References

Burrows, J.F. (1987). Word patterns and story shapes: The statistical analysis of narrative style, *Literary and Linguistic Computing*, 2, 61-70.

Burrows, J.F. (1989). 'A vision' as a revision, *Eighteenth Century Studies,* 22(4), 551-565.

Burrows, J.F. (1992). Computers and the study of literature. In C. Butler (ed.), *Computers and Written Texts: An Applied Perspective*. Oxford: Blackwell, 167-204.

Burrows, J.F. and Craig, H.D. (1994). Lyrical drama and the "turbid mountebanks": Styles of dialogue in romantic and renaissance tragedy, *Computers and the Humanities*, 28(2), 63-86.

Burrows, J. F. and Hassal, A. J. (1988). Anna Boleyn and the authenticity of Fielding's feminine narratives, *Eighteenth Century Studies*, 21(4), 427-453.

Canter, D. V. (1992). An evaluation of the "Cusum" stylistic analysis of confessions, *Expert Evidence*, 1(3), 93-99.

Diederich, J., Kindermann, J., Leopold, E. and Paass, G. (2003). Authorship Attribution with Support Vector Machines, *Applied Intelligence*, 19(1), 109-123.

Foster, D. W. (1989). *'Elegy' by W.S.: A study in attribution.* Cranbury, NJ: Associated University Presses.

Hardcastle, R.A. (1997). CUSUM: a credible method for the determination of authorship?, *Science & Justice,* 37(2), 129-138 (doi: 10.1016/s1355-0306(97)72158-0).

Joachims, T. (1998). Text categorization with Support Vector Machines: Learning with many relevant features. In Nédellec C. and Rouveirol C. (eds.), *Proceedings of the 10th European Conference on Machine Learning, 21-24 April 1998, Dorint-Parkhotel, Chemnitz, Germany*. Berlin: Springer, 137-142.

Kuhn, M., Wing, J., Weston, S., Williams, A., Keefer, C. and Engelhardt, A. (2012). caret: Classification and Regression Training: R package version 5.15-023.

Mikros, G. K. (2013). Authorship Attribution and Gender Identification in Greek Blogs. In Obradović, I., Kelih E. and Köhler R. (eds.), *Selected papers of the VIIIth International Conference on Quantitative Linguistics (QUALICO) in Belgrade, Serbia, April 16-19, 2012*. Belgrade: Academic Mind, 21-32.

Mikros, G. K. and Perifanos, K. (2011). Authorship Identification in Large Email Collections: Experiments Using Features that Belong to Different Linguistic Levels – Notebook for PAN at CLEF 2011. In Petras, V., Forner, P. and Clough, P. D. (eds.), *CLEF 2011 Labs and Workshop, Notebook Papers, 19-22 September 2011, Amsterdam, The Netherlands* (Vol. 1177). Amsterdam, The Netherlands: CEUR-WS.org, 1-6.

Mikros, G. K. and Perifanos, K. (2013). Authorship attribution in Greek tweets using multilevel author's n-gram profiles. In Hovy, E., Markman, V., Martell, C. H. and Uthus, D. (eds.), *Papers from the 2013 AAAI Spring Symposium "Analyzing Microtext", 25-27 March 2013, Stanford, California*. Palo Alto, California: AAAI Press, 17-23.

Mosteller, F. and Wallace, D. L. (1984). *Applied bayesian and classical inference. The case of The Federalist Papers* (2nd ed.). New York: Springer-Verlag.

Reddy, T. R., Vardhan, B. V. and Reddy, P. V. (2016). A Survey on Authorship Profiling Techniques, *International Journal of Applied Engineering Research*, 11(5), 3092-3102.

Sanford, A. J., Aked, J. P., Moxey, L. M. and Mullin, J. (1994). A critical examination of assumptions underlying the Cusum technique of forensic linguistics, *Forensic Linguistics*, 1(2), 151-168.

Smith, M. W. A. (1990). Attribution by statistics: a critique of four recent studies, *Revue Informatique et Statistique dans les Sciences humaines*, 26, 233-251.

Stamatatos, E. (2009). A survey of modern authorship attribution methods, *Journal of the American Society for Information Science and Technology*, 60(3), 538-556 (doi: 10.1002/asi.21001).

Tuzzi, A. and Cortelazzo, M. A. (2018). What is Elena Ferrante? A comparative analysis of a secretive bestselling Italian writer, *Digital Scholarship in the Humanities*, (online first fqx066-fqx066. doi: 10.1093/llc/fqx066).

Vapnik, V. (1995). *The nature of statistical learning theory*. New York: Springer-Verlag.

The Brilliant Friend(s) of Elena Ferrante: A Lexicometrical Comparison between Elena Ferrante's Books and 39 Contemporary Italian Writers

Pierre Ratinaud
Université of Toulouse 2 – Jean Jaurès

Abstract

This article shows results of the comparison of 7 books of Elena Ferrante with 140 other books writing by contemporary Italian authors. This comparison is exclusively built from textual statistics with the software IRaMuTeQ (Ratinaud, 2014). Classical methods are used to study the relationships between words written by all these authors: correspondence analysis on complete lexical tables (Lebart and Salem, 1994), distance computation with the Labbe index (Labbé and Monière, 2000) and hierarchical clustering with the Reinert methods (Reinert, 1983; Ratinaud and Marchand, 2012). These methods are computed on different partitions of the corpus. First, all books of each author are considered as a whole, then each book is studied as a unit, and finally we use clusters computed on each book. All these results converge to the same finding: words used in Elena Ferrante's books are closer to the ones used by Starnone than to any other authors in the sample.

Introduction

Statistical analysis of textual data is a practice performed in many fields of research. In human sciences, the methodologies grouped under this name are used in fact in all disciplines, contributing to very diverse issues. The corpora analyzed show a large variability in terms of genre and size. All researchers using these methods are aware of the time required to build this type of corpus. First, I would like to thank Arjuna Tuzzi and Michele Cortelazzo for the compi-

lation of texts that I am about to analyze. This remark allows me also to clarify that, a priori, I would never have considered the "Ferrante mystery" since I do not work on literature. Usually, my work focuses on the study of social and professional representations (Moscovici, 1961; Piaser, 1999; Ratinaud and Lac, 2011; Ratinaud and Marchand, 2015) and their dynamics that lead me to build and to analyze corpora from interviews, newspapers articles or data from socio-digital networks. This is to attest that I am not a specialist in literature studies. Furthermore, the field of authorship attribution is not my specialty and I would not venture to follow this way. It is also to be noted that I do not speak Italian.

These remarks make it possible to emphasize one of the interests, within the framework of a scientific approach, of the methods that I propose to use to analyze a corpus of 147 novels from 40 different Italian writers. These methods are based on computational and statistical processing of texts which are independent, before the interpretation of the results, of the knowledge of the researcher on these texts. All these analyses rely on a count of words, comparisons of the proportion of word frequencies between texts and counts of co-occurrences of words in texts or portions of text. The objective of these techniques is to study the closeness or distance between texts or authors based on the lexicon they mobilize. All the manipulations required to produce these results are independent of the meaning of the words manipulated or the way they are represented. A word-for-word translation of this corpus into any other language would produce nearly the same results. Actually, even a real translation could produce the same results (Rybicki, 2012). All these analyses are carried out by the free software IRaMuTeQ (Ratinaud, 2014).

After the description of the corpus as it is formalized in the tool, we follow an approach that involves different levels of granularity, from the most general to the finest. First, we consider the authors as a unit, grouping all of their works into one collection, then we use the book as a unit and finally, we manipulate the "lexical worlds" determined on each book. The analysis of these different partitions are mainly based on correspondence analysis of full lexical tables (Lebart & Salem, 1994) and of lexical distances computation with the Labbé's index (Labbé & Monière, 2000; Labbé & Labbé, 2001; 2013).

Description of the corpus

The corpus is slightly different from the one used by my colleagues. In comparison with the original sample, 3 books are missing in this collection. These are Baricco 1993, Faletti 2006 and Milone 2015. The corpus is made up of 147 books written by 40 different authors. Table 1 summarizes this corpus with lexicometric indicators. These indicators illustrate the segmentation of the corpus

in the software. They are the result of a series of interventions on the text that are not neutral, but applied systematically to all works in the collection.

Table 1: Description of the corpus

texts	147	
tokens	9405562	
	types	lemmas
	168042	103768
hapax	64286	46631
% hapax (types)	38.2 %	44.9 %
% hapax (tokens)	0.68 %	0.5 %

The corpus is therefore composed of about 9.5 million of tokens. 168042 different words are present, of which 64286 (38.2%) are hapax[1]. Lemmatization, which consists in reducing the verbs to the infinitive, the adjectives to the singular masculine and nouns to singular, holds 103768 lemmas.

Authors as a unit

For this first series of analyses, we use the authors as a unit. For example, for Elena Ferrante, this analysis considers the 7 novels of the collection as a whole, and the entire lexicon of these 7 books is processed as one big book. Figure 1 shows the result of a correspondence analysis carried out on the full lexical table which partitions this corpus according to the authors. This table is a simple contingency table with authors in column and words in row. The cells in the table contain the frequency of each word in each author. The analysis allows to project on a 2 dimensional plane the relations between the authors on the basis of their lexical co-occurrences, i.e. their tendency to use the same words (or not to use the same words).

The factorial plan produced by this first analysis shows an opposition on the first factor (the horizontal factor) between Ferrante and Starnone to the left of this factor, and all the other authors, in the center or on the right of this factor. Ferrante and Starnone are the two authors who "contribute" the most to the construction of this factor. This result brings two comments. It seems on the one hand, that the lexicon mobilized by Ferrante and Starnone presents some proximities. On the other hand, this phenomenon is accentuated by the fact that these two authors present lexical features which are not present in the other authors of the collection. Although this type of plan can sometimes be quite com-

[1]An hapax is word that only appears one time in a corpus.

plex to interpret (this plan is a 2-dimensional representation of an analysis that extracted 39 factors), the effect we observe here is particularly easy to detect.

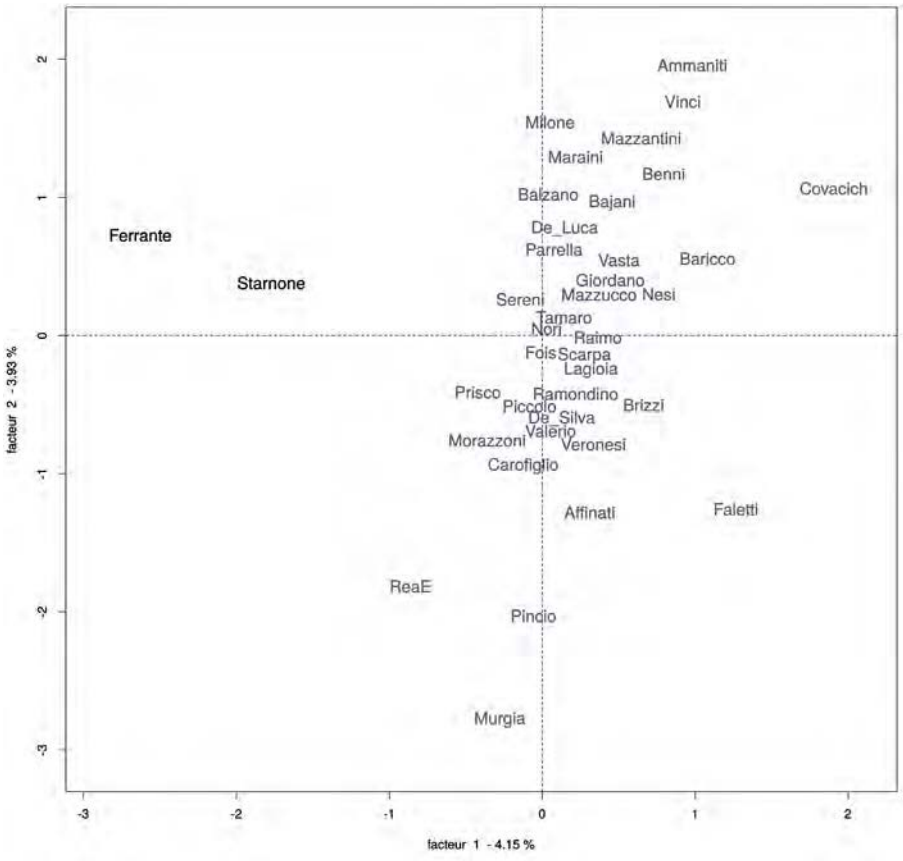

Figure 1: Correspondence analysis on the full lexical table of authors

As with the other segmentations, we have submitted this same partition to a second analysis based on the calculation of the lexical distance between the authors with Labbé's index (Labbé & Monière, 2000; Labbé & Labbé, 2001). This index allows to assess to what extent two texts are close to or distant from the point of view of the lexicon that compose them.

Labbé's distance

The analysis is done in two stages: first, the distance between each pair of authors is calculated with Labbé's index. This index summarizes the differences

of frequency of each words of texts. It produces a score between 0 and 1 where 0 reports two identical texts and 1 two texts without any commun words. The computation of this score for each pair of authors generates a distance matrix (square matrix with authors in column and row). In a second step, we compute a divisive hierarchical clustering on this matrix to simplify the interpretation (we use Ward's method here).

Figure 2 is a tree representation of the classification led on the distance matrix. It shows that the distance between Ferrante and Starnone is the lowest of all the distances calculated in the matrix. In other words, this analysis allows us to conclude that the lexicon used by Ferrante, in all of her work, is closer to the lexicon used by Starnone than to the lexicon of any other authors in this collection. We can also see some similarities between the first both analyzes. For example, we note the group formed by Rea, Murgia and Pincio, both at the bottom of the correspondence analysis (Figure 1) and at the bottom left of the tree (Figure 2).

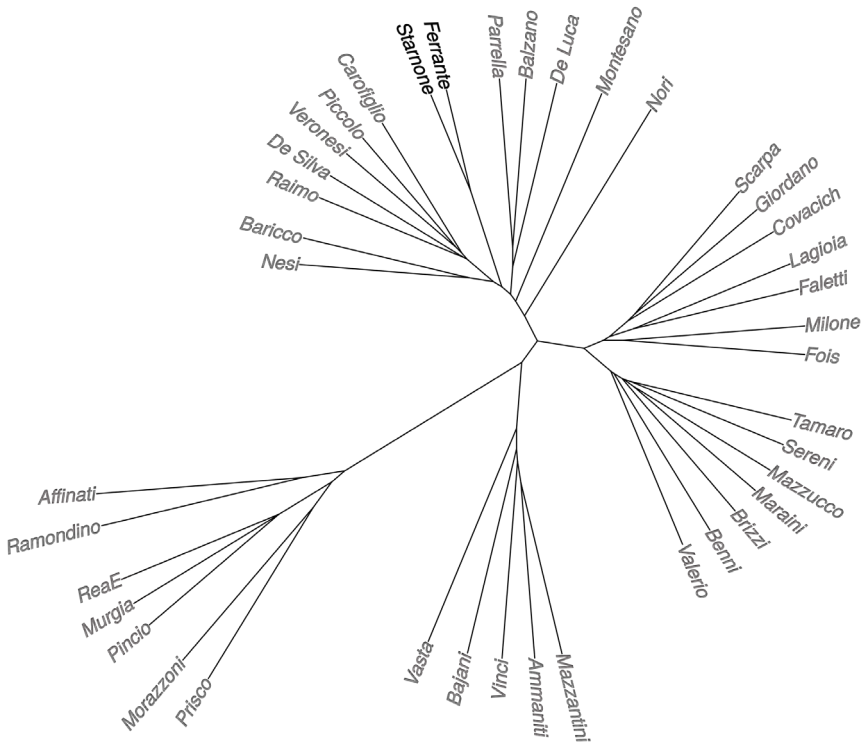

Figure 2: Clustering on Labbé's distances of authors

The proximity between Starnone and Ferrante in the correspondence analysis can easily be found on other dimensions present in the corpus. For example, if we look at the birthdate or at the birthplace of the authors, we see that Starnone is the only one of the collection to be born in 1943 and to be born in Saviano. In consequence, we can practically reproduce the Figure 1 on these variables.

In Figures 3 and 4, the symbol "???" refers respectively to Ferrante's birthplace or birthdate. We see that each of these analyses opposes the variables related to Starnone and Ferrante to those of all the other authors. In these analysis, the contingency table at the origin of the calculations is practically identical to the one used for Figure 1 (in this table, the column representing Ferrante is identical to the preceding table and the columns 1943 or Saviano are identical to the Starnone column), which explains that the process leads to the same results.

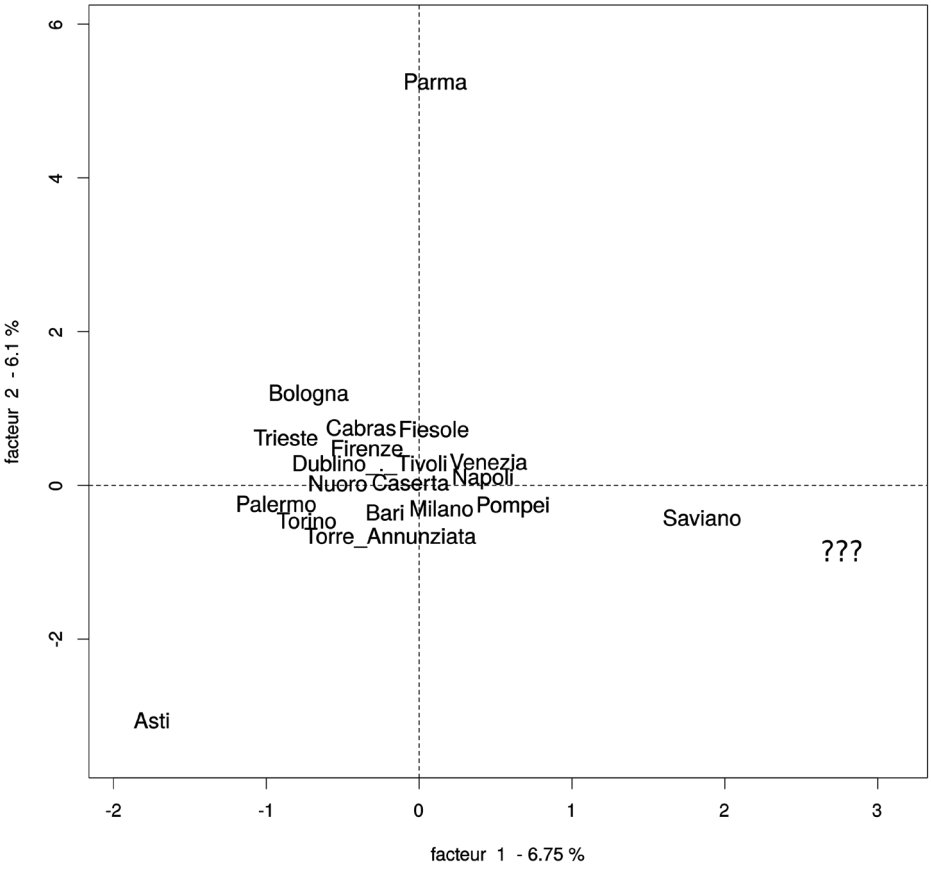

Figure 3: Correspondance analysis on the full lexical table of birthplaces. «???» represents Elena Ferrante

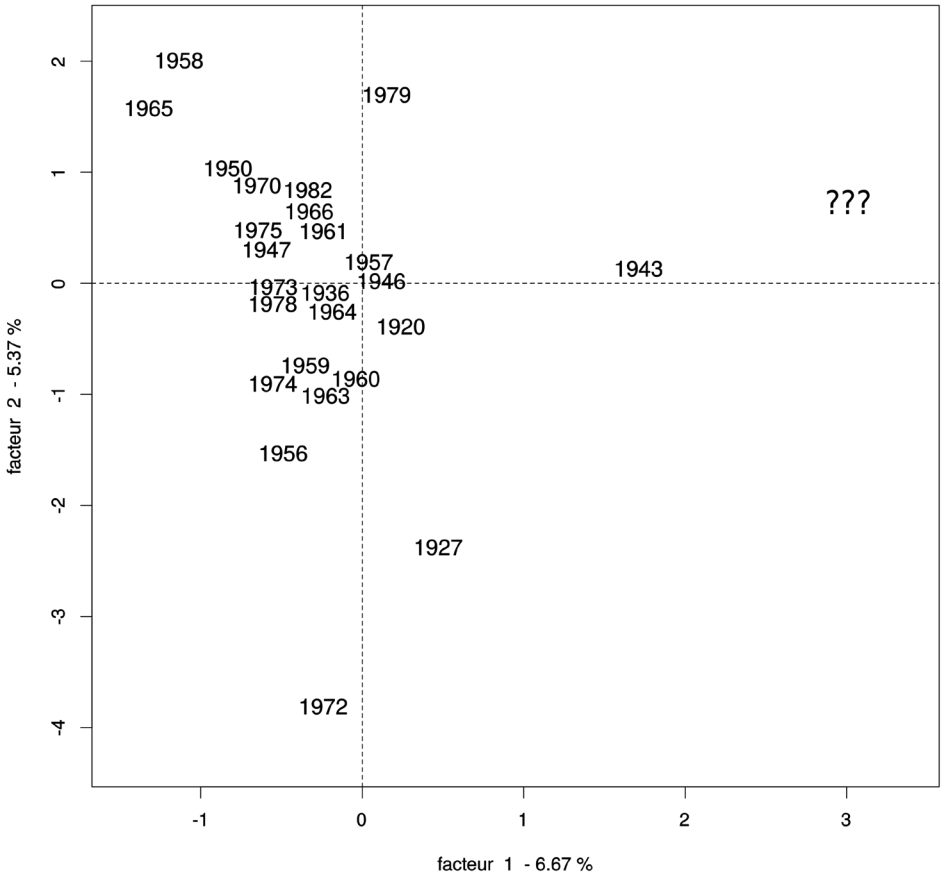

Figure 4: Correspondence analysis on the full lexical table of birthdates. «???» represents Elena Ferrante.

We can see from the authors' regions of origin analysis (Figure 5) that the fact that Starnone shares its region of birth (Campania) with other authors tends to "move" this variable towards the center, even if it continues to be attracted by the variable marking the texts of Ferrante ("???").

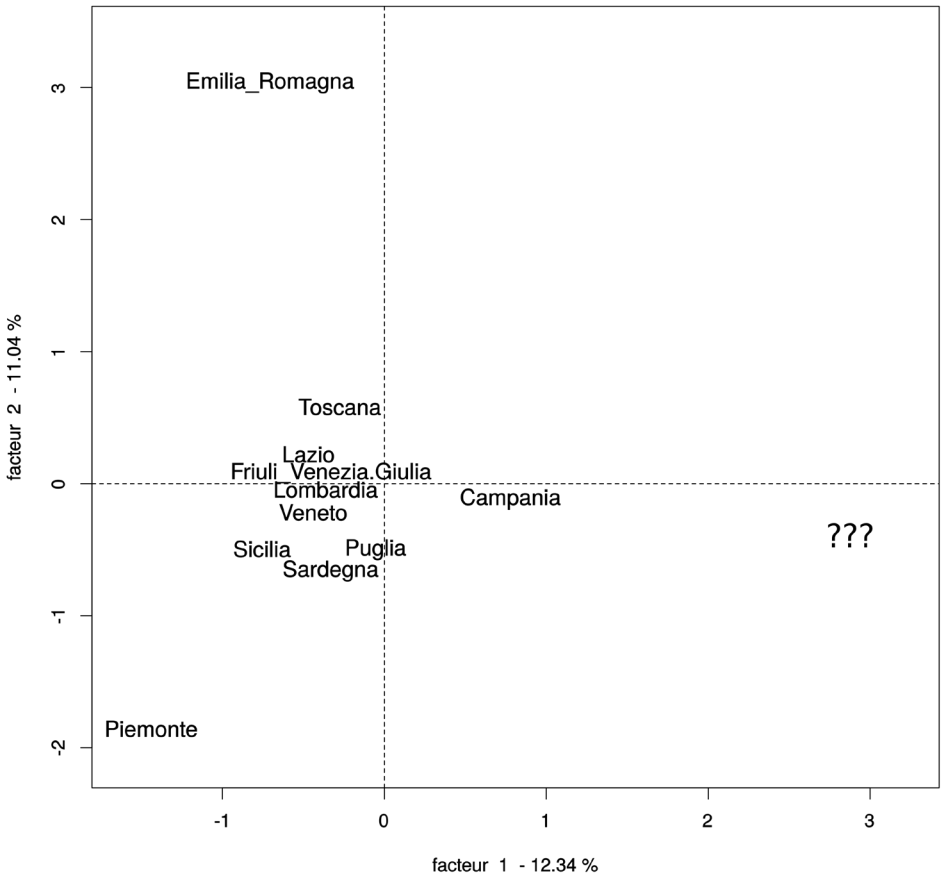

Figure 5: Correspondence analysis on the full lexical table of region of origin. «???» represents Elena Ferrante.

Novels as a unit

These first analyses can be made more specific by lowering the level of granularity of the manipulated textual units. In the following analyses, we use the novel as a variable to partition the texts in our corpus. The correspondence analysis presented in the Figure 6 is thus operated on a contingency table which presents the novels in column and the words in row.

We find here the lexical proximity between the works of Starnone and Ferrante, but subgroups of novels seem to appear. We can see, for example, Ferrante's quadrilogy takes a special place. These 4 books are close to each oth-

er on this factorial plane because they share a very close lexicon. The lexical proximity is here amplified by the unity of character and place of this series of work. They largely explain the polarity of Ferrante variables on each analysis. We also note that the Starnone's novel written in 1987 seems to differentiate itself. Closer to the center, Ferrante's other novels come closer to Starnone's works, but they are also close to other authors. We note here the limit of interpretation of correspondence analysis. It does not allow us to say whether the Ferrante written in 1992, 2002 and 2006 are closer to the novels of Starnone than to novels of other authors. The use of Labbé's lexical distance is more accurate when the number of variables increases. We have reproduced the same analysis as previously (calculation of the distance matrix and tree representation of the classification) on the novels. Figure 7 reports the results.

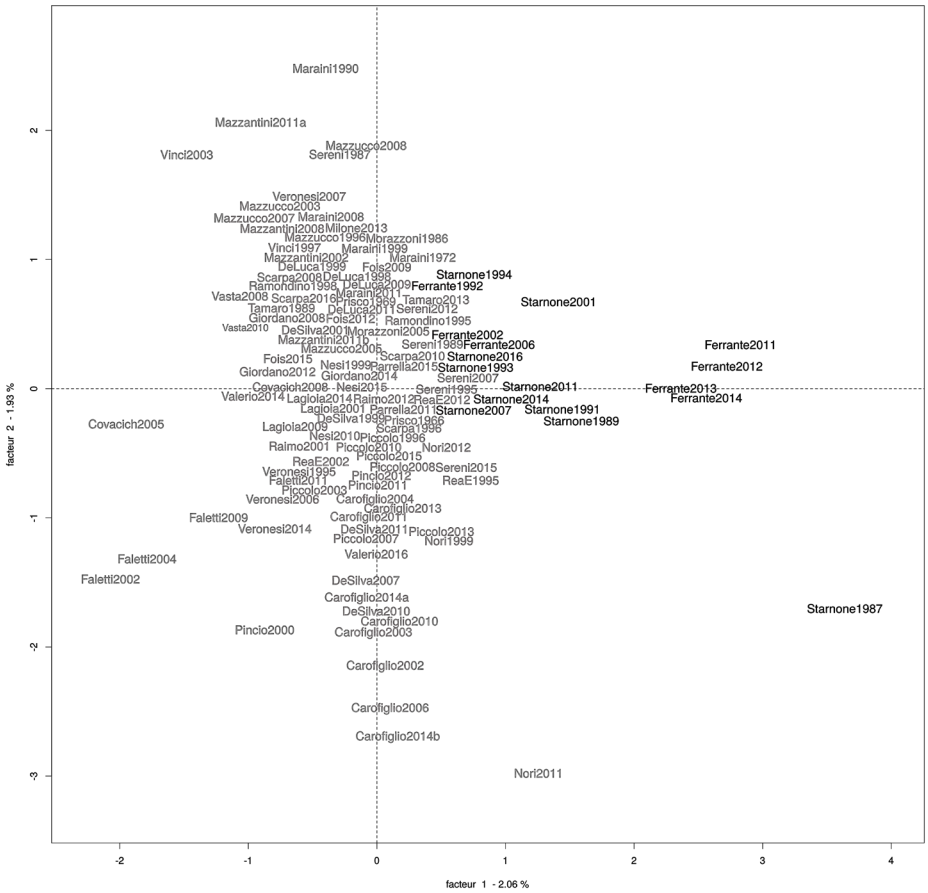

Figure 6: Correspondence analysis on the full lexical table of novels

This analysis allows several comments. It seems quite obvious here that the novels are first organized by author. When genre is circumscribed, it is the authors of the texts who have the most weight on the lexicon mobilized (Brunet, 2016a; 2016b). We can therefore see that, with rare exceptions, all the novels of all authors are found in the same subgroups in this analysis. Significantly, these exceptions include the works of Ferrante and Starnone. Even if subgroups are formed (for example the subgroups of the novels of the Ferrante's quadrilogy or the Starnone's novels of 1987, 1989 and 1991), the works of Starnone and Ferrante are mixed. This tells us that the lexical proximity that we observe between these both authors is found more particularly in some of their novels. Moreover, none of their novels is far from others in terms of lexicon. The lexical congruence between these both authors is therefore almost constant over time, regardless of the evolution of their works.

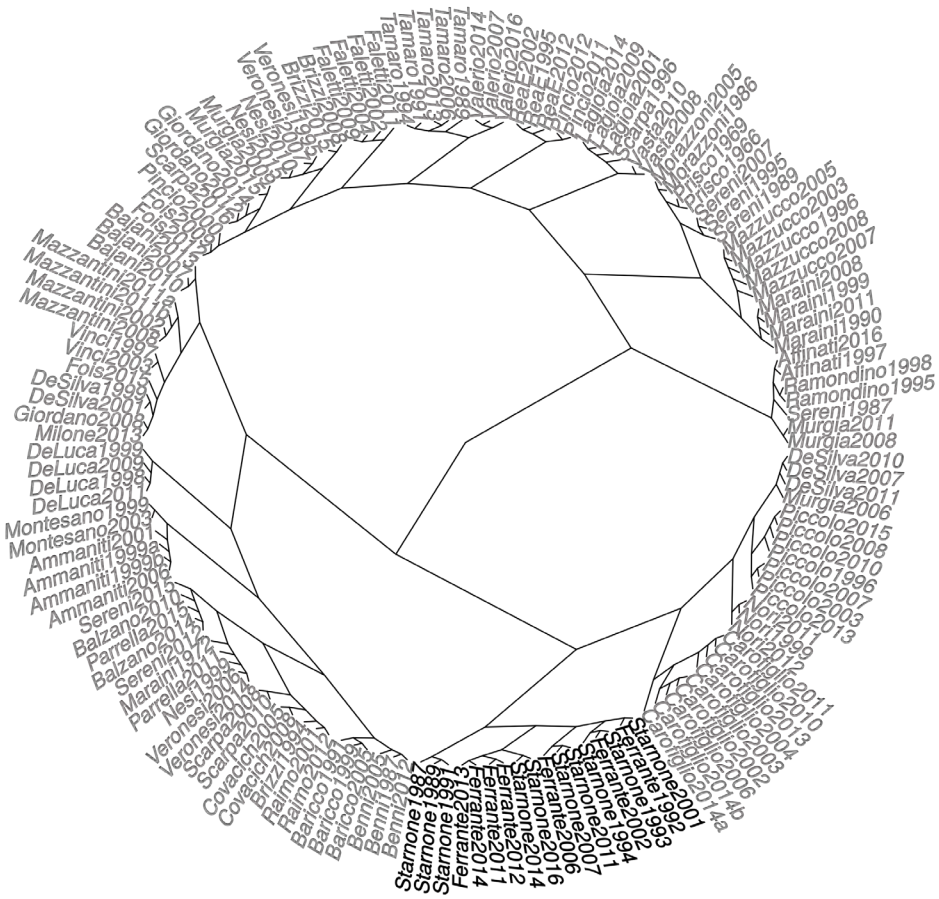

Figure 7: Clustering on Labbé's distances of novels

Topics as a unit

In order to complete and replicate the previous results on another dimension of this collection, we submit this corpus to an experiment. This consists in applying the preceding analysis to another granularity extracted from the novels. Each of the 147 novels is processed through a Reinert type analysis (Reinert, 1983; 1990; Ratinaud and Marchand, 2012). This analysis allows to determine topics in corpora. It is based on a partition of each book into segments of text of a size corresponding to roughly to large sentences. Each of the texts is thus partitioned into text segments of about forty tokens. The analysis starts with a matrix that crosses these segments of text and "full forms" (nouns, verbs, adjectives and adverbs). Then, it proceeds to a divisive hierarchical clustering of the segments (thus the rows of the matrix) which is based on a series of bi-partitions obtained from a correspondence analysis. The objective is to group the text segments into sets called "clusters" on a criterion of lexical co-occurrence. In other words, the analysis produces sets of segments that tend to contain the same words and therefore to refer to the same topic.

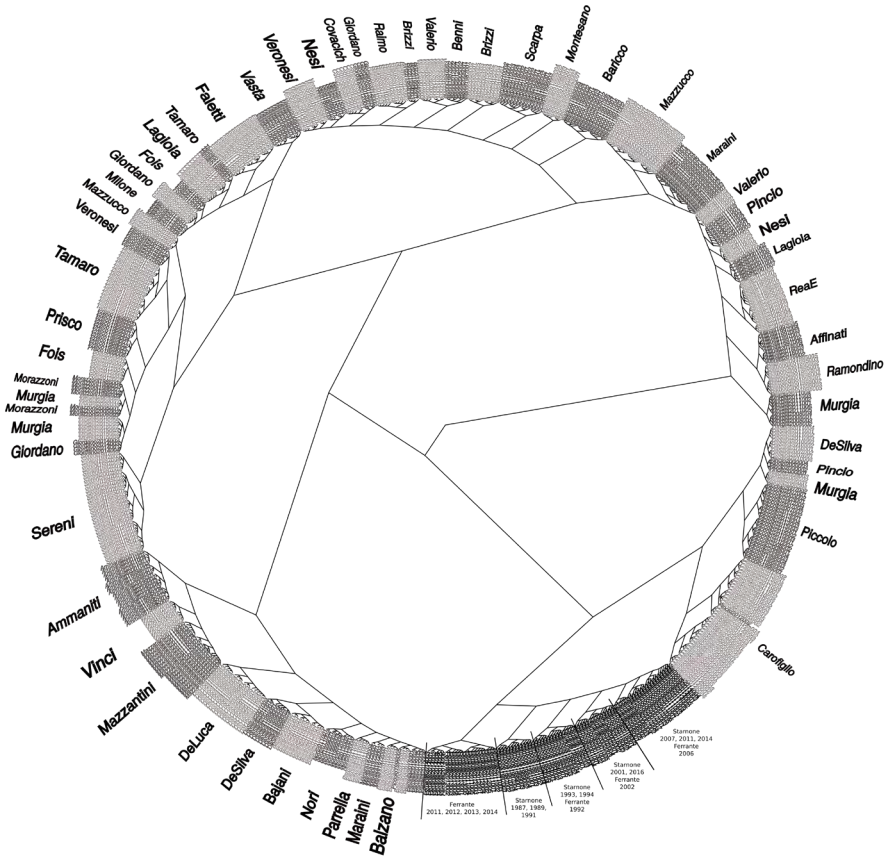

Figure 8: Clustering on Labbé's distances of clusters

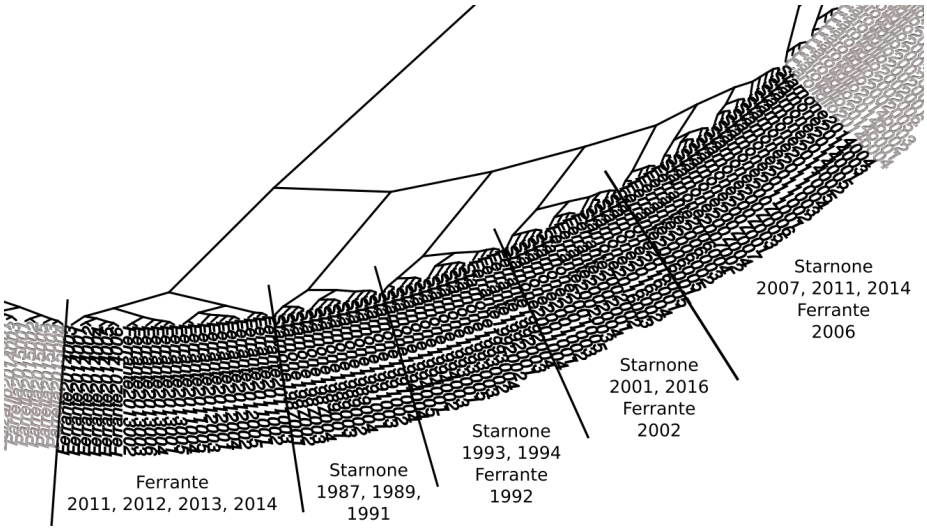

Figure 9: Zoom on the bottom right of Figure 8

We only present here the classification on Labbé's distance matrix. This analysis is finally very close to the analysis on the novels. With rare exceptions, all the clusters from each novels are grouped together and the novels of the same author are grouped as in the previous analysis. The positioning of the works of Ferrante and Starnone is also similar to the previous analysis. From left to right in Figure 9, we find the 4 novels of the quadrilogy together and then the first three novels of Starnone. The other sets are a mixed of the works of Starnone and Ferrante and these groups follow the chronology of the publications. Thus, the Ferrante of 1992 is placed with the Starnone of 1993 and 1994; the Ferrante of 2002 with the Starnone of 2000 and 2016; the Ferrante of 2006 with the Starnone of 2007, 2011 and 2014.

Conclusion

All the analyses that we have presented converge. Then we note that one of the reasons for this convergence is the unity of the corpus of the analyses. If these analyses converge, it is also because they are made on the same single corpus. Even if it has undergone different divisions (by authors, by novels or by clusters of novel), the corpus at the origin of the analyses is always the same. This is the main limitation of these studies: their interest essentially depends on the completeness of the original corpus.

In this set of texts and authors, all these analyses lead to a finding that seems difficult to refute: in this corpus, the lexicon mobilized by Elena Ferrante in the 7 analyzed novels is closer to the lexicon mobilized in the novels of Starnone than that mobilized in the works of all other authors. This is a statistical observation that any analyst would make if he or she uses the same collection with the same statistical tools. Analyses on lower granularities make it possible to clarify this observation: the proximity between the lexicon of Starnone and Ferrante is present throughout the career of these authors, as evidenced by the results obtained on the novels and on the clusters made on novels.

But these analyses do not allow us to deal with the hypothesis that we must formulate: there may be, in Italian contemporary literature, an author who presents a lexicon closer to that of Ferrante than is that of Starnone. It is unfortunately not possible to test this hypothesis satisfactorily, because it is not really possible to build this kind of "complete" corpus. It seems to me that the corpus proposed by Arjuna Tuzzi and Michele Cortelazzo is the best approach we have for now.

References

Brunet, E. (2016a). *Ce qui compte*. Paris: Honoré Champion.

Brunet, E. (2016b). *Tous comptes faits*. Paris: Honoré Champion.

Labbé, D. and Monière, D. (2000). La connexion intertextuelle. Application au discours gouvernemental québécois. In Rajman, M. and Chappelier, J.-C. (eds.), *Actes des 5èmes Journées Internationales d'Analyse statistique des Données Textuelles*. Lausanne: EPLF, 85-94.

Labbé, C. and Labbé, D. (2001). Inter-textual Distance and Autorship Attribution. Corneille and Molière, *Journal of Quantitative Linguistics*, 8(3), 213-231.

Labbé, C. and Labbé, D. (2003). La distance intertextuelle. *Corpus*, 2 (consulté à l'adresse https://corpus.revues.org/31?lang=en).

Lebart, L. and Salem, A. (1994). *Statistiques textuelles*. Paris: Dunod.

Moscovici, S. (1961). *La psychanalyse, son image et son public: étude sur la représentation sociale de la psychanalyse*. Paris: Presses universitaires de France.

Piaser, A. (1999). *Représentations professionnelles à l'école. Particularités selon le statut: enseignant, inspecteur.* Thèse de doctorat en Sciences de l'éducation. Université de Toulouse Le Mirail, Toulouse.

Ratinaud, P. (2014). *IRaMuTeQ: Interface de R pour les Analyses Multidimensionnelles de Textes et de Questionnaires* [software, Version 0.7 alpha 2]. Retrieved from http://www.iramuteq.org

Ratinaud, P. and Lac, M. (2011). Understanding professionalization as a representational process. In Chaïb, M., Danermark, B. and Selander, S. (eds.),

Education, Professionalization and Social Representations – On the Transformation of Social Knowledge. New-York: Routledge, 5567.

Ratinaud, P. and Marchand, P. (2012). Application de la méthode ALCESTE à de "gros" corpus et stabilité des "mondes lexicaux": analyse du "CableGate" avec IRaMuTeQ. In *Actes des 11eme Journées internationales d'Analyse statistique des Données Textuelles*, 835-844.

Ratinaud, P. and Marchand, P. (2015). Des mondes lexicaux aux représentations sociales. Une première approche des thématiques dans les débats à l'Assemblée nationale (1998-2014), *Mots. Les langages du politique*, 108, 57-77.

Rybicki, J. (2012). The Great Mystery of the (Almost) Invisible Translator: Stylometry in Translation. In Oakley, M. and Ji, M. (eds.), *Quantitative Methods in Corpus-Based Translation Studies*. Amsterdam: John Benjamins, 231-248.

Reinert, M. (1983). Une méthode de classification descendante hiérarchique: application à l'analyse lexicale par contexte. *Les cahiers de l'analyse des données*, 8(2), 187-198.

Reinert, M. (1990). ALCESTE. Une méthodologie d'analyse des données textuelles et une application: Aurélia de Gérard de Nerval. *Bulletin de méthodologie sociologique*, 26, 24-54.

Partners in Life, Partners in Crime?

Jan Rybicki
Jagiellonian University of Kraków, Poland

Abstract

A series of stylometric tests for authorship, based on Burrows's Delta procedure, which compares usage of most frequent words, was run on a corpus of novels by contemporary Italian writers, supplemented with translations by Anita Raja, recently the main suspect for being Elena Ferrante. Rather than to Raja, the tests point overwhelmingly to her husband, the writer Domenico Starnone.

Introduction

Of course, there is no crime in the story of Elena Ferrante. In literature, it is quite all right to pretend to be someone else, especially someone who does not really exist – as opposed to pretending to be someone who does, or to take the works of existing persons and pass them off as one's own. Jane Austen originally published as "A Lady," the Brontë sisters as the Bell brothers. Octave Mirbeau first earned his reputation *en négritude*. J.K Rowling hid so well behind the Galbraith persona that "his" books really started selling after the discovery, or the leak, that *The Cuckoo's Calling* shared its author with the Potter series. Aleksander Głowacki wrote his novels as Bolesław Prus and Samuel Langhorne Clemens as Mark Twain, and in both cases, and in many other, the preference is to use the pseudonym rather than the real name.

The Ferrante case is different because the Ferrante pretence is now being kept up against all odds and against all stylometric and financial evidence presented, among other, in this collection of experiments. One of the two main suspects says very deliberately, "I. Am. Not. Elena. Ferrante" (van der Ploeg,

2017); but that, too, is not a crime. And perhaps it is better that way: things would become a little less interesting with an avowal of authorship; as it is, there remains that titillating tension between – to steal the opposition from Adam Mickiewicz's Romantic manifesto of 1821, translated so freely and so surprisingly by W.H. Auden – the *love* of the Elena Ferrante persona or the *faith* shared by a great number of her serious readers that the ideas and situations definitely point to a female author and definitely exclude the possibility of a male hand; and the *lenses* of multivariate analysis to the writers' patterns of usage of most-frequent words. Recently, the stakes have even been raised in this game when "Elena Ferrante" seems to openly challenge stylometric sleuths by her new (translated) column in *The Guardian*…

Method and Material.

Still, my title is not the only thing that makes it tempting to apply a "criminal" metaphor: all kinds of quantitative approaches to authorship attribution have at one time or another mentioned dactyloscopy. Well in the previous century, Kenny speaks of a "stylistic fingerprint,… a combination perhaps of very humble features such as the frequency of 'such as' — no less unique to him that a bodily fingerprint is. Being a trivial and humble feature of style would be no objection to its use for identification purposes: the whorls and loops at the ends of our fingers are not valuable or striking parts of our bodily appearance" (Kenny 1982, p. 12). A reviewer of one of my papers has recently shocked me by saying that there is no epistemological connection between the occurrences of most frequent words and some of the general literary assumptions we stylometrists have been reading into multivariate analyses of those frequencies at least since Burrows's "Computation" of the "Style" of Jane Austen (1987). Statistics-based authorship attribution usually fares slightly better than those outlandish distant readings and macroanalyses, having acquired a degree of respect from (sometimes highly unrespectable) squabbles over who wrote what with or without Shakespeare, but it is usually believed in proportion to the relative distance of a scholar's desk between the linguistics and the literary departments of his or her institution. Even if it is believed, both literary and linguistic scholars express dismay at the fact that authorship attribution only needs several hundred very frequent (and thus very "unmeaningful") words to make its determinations, rather than some extraordinary turns of phrase (literary scholars would say), or part-of-speech n-grams (linguists complain). To be quite honest, this phenomenon even puzzles some stylometrists – including this one.

My search for the hand that held the pen (or, perhaps, that punched the keyboard) that produced *L'amica geniale* and the other bestsellers of the anony-

mous Italian writer was made very simple by two things. First, the investigation by Claudio Gatti into the financial rather than the stylistic traces of Elena Ferrante, which led him to the Raja-Starnone household (2016); second, the creation, by the editors of this volume, of a 150-strong corpus of texts by the usual, the possible (and the less possible) suspects and obviously containing the entire output of Ferrante (Tuzzi and Cortelazzo, 2018). To this collection, I added 16 Italian translations of works by Christa Wolf: 14 by Anita Raja and one each by two other translators. To obtain an even broader perspective, I used my own collection of some 1200 Italian dramas, fiction and epic poetry I have scraped from Liber Liber's invaluable *Progetto Manuzio*.

All these texts were in plain UTF-8 text format, which is the preferred input for *stylo*, a package (Eder et al., 2016) written for the R statistical programming environment (R Core Team, 2014). It processed the texts by dividing them into word tokens and calculating the occurrences of their word-types in the entire collection to establish the ranking list of up to 5000 most frequent word-types. In the next stage, the frequency of the word-types was counted in each individual text, and relative frequencies were calculated in reference to each individual text lengths. It is those series of relative frequencies that were compared to establish a measure of distance between each pair of texts; in this case, cosine similarity, recently shown to be highly accurate in authorship attribution (Evert et al., 2017) was used. Cosine Delta ($\Delta\angle$), as it often referred to, for two texts (T and T_1), measures the angle α (the greater the angle, the greater the distance) between the two texts, and this angle is given by the formula,

$$\cos \propto = \frac{\sum_{i=1}^{n_s} x_i y_i}{\sqrt{(\sum_{i=1}^{n_s} x_i^2)}\sqrt{(\sum_{i=1}^{n_s} y_i^2)}},$$

where x = z(T) and y = z(T_1), and z(T) is the value of z-score of word frequency in text T, calculated according to the usual formula,

$$z(T) = \frac{f_s(T) - \mu_s}{\sigma_s},$$

where $f_s(T)$, in turn, is the raw frequency of a given word s in text T, μ_s the average frequency of word s in the set of texts to which T belongs, and σ_s is the standard deviation of the frequency of word s in that same set of texts (Smith and Aldridge, 2011).

This produces a whole matrix of distances; and while this might be enough to find pairs of similar texts with the least distance between them – and it is stylometry's axiom that these are usually written by the same hand – it is safer

to process such a matrix with a multivariate statistical method. In this case, Ward's hierarchical clustering was used at multiple iterations of most-frequent-word frequency list (from 100 to 2000 with an increment of 100) to produce a "consensus tree" that shows the most consistent nearest neighbours amongst the texts studied. This consensus approach is useful for relatively small collections of texts, but when their number increases, diagrams usually become too cluttered. This is probably what prompted the use of network analysis, starting at least with Jockers's *Macroanalysis* (2013).

The procedure used in this study has been described by Eder (2017) and is based on attributing "weights" of different values (5, 3 and 1 in this case) to, respectively, any pair of nearest-neighbours, next-to-nearest-neighbours and next-to-next-to-nearest neighbours. The software used to process the output from *stylo* is Gephi (Bastian et al, 2009); in this study, it applies a gravitational algorithm, Force Atlas 2, which applies multidirectional pull to weigh the system of stronger and weaker connections until a balance is reached which best reflects the overall pattern of similarity and difference between pairs of texts (Jacomy et al., 2008).

Results

The first figure (Fig. 1) cannot be seriously treated as a determination of the authorship of Ferrante's novels, since it presents almost 1500 texts written in Italian, in a variety of genres, starting with Dante and ending in the present. At the same time, it might be a useful illustration of some very usual phenomena that come to the fore in this sort of visualization. It quite typical, then, for a major split to occur between the main literary genres. This is visible in the way in which the purely dialogic dramatic/operatic texts, represented here among others by Goldoni, Metastasio and Italian translations of Shakespeare extend to the left, while narrative/dialogic prose and epic poetry evolves to the right. It is quite logical that Manzoni, represented by his novels as well as by his plays, is suspended in the middle. Also, "evolution" is a very appropriate word here, as this part of the graph exhibits strong chronological ordering from bottom-centre to top-left. The separate clusters for Pirandello, Deledda and Salgari are a very normal phenomenon associated with authors who produce a large body of stylometrically-uniform work.

The main protagonists of this study were also included in this diagram, and they are visible, predictably, among the rest of late 20th/early 21st century fiction in the top-right quadrant of the previous picture (Fig. 2). The black patch of Starnone is quite close to the dark-grey cluster of Ferrante, while Raja's translations of Christa Wolf (medium grey) are spread further down against a background of

light-grey modern literature. Obviously, this is not enough to make any statements on authorship, since authors of any texts within that background could aspire to being Ferrante; but this can already serve as some indication of where to look for her: in the above-mentioned Tuzzi/Cortelazzo corpus of 150 contemporary novels, with the added "bonus" of the Raja translations.

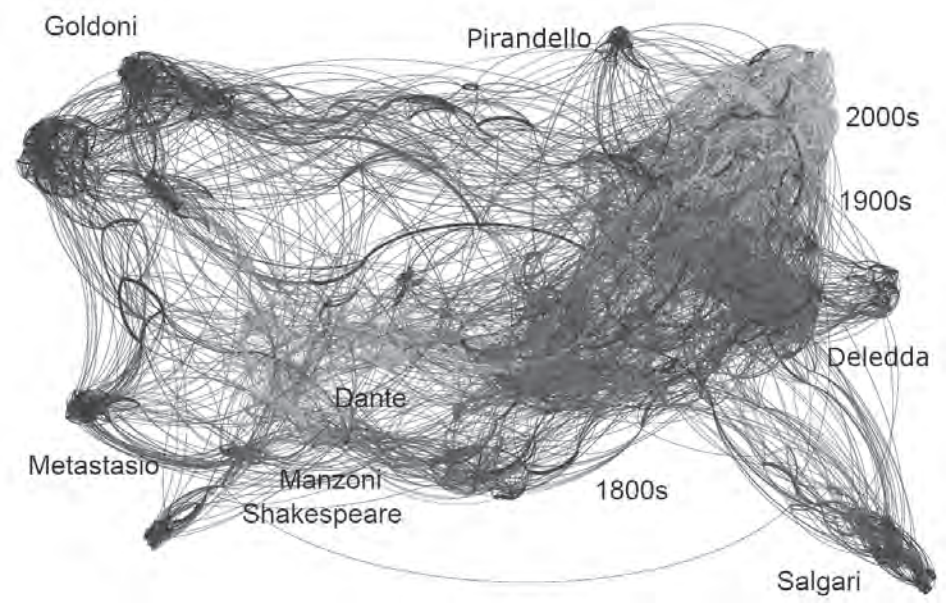

Figure 1. Network analysis of ca. 1500 texts of Italian literature.

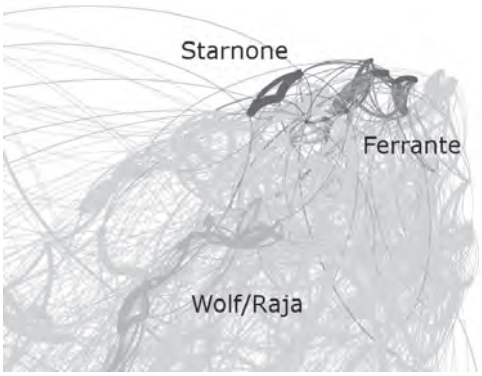

Figure 2. Texts by Ferrante, Starnone and Raja's translations of Wolf in a fragment of the large Italian literature network.

This is done in Figure 3, and things become quite clear. A separate cluster of black and dark grey nodes and edges (points and linkages) of the network grows upward from the main light-grey body of the corpus. Texts by Starnone are in black; texts signed "Ferrante" are dark grey. It is not just that they are close to each other; more importantly, they become peripheral *together*. Thus the above-mentioned phenomenon of outlying clusters of coherently similar large bodies of single-author work – which is often deplored by stylometrists as it might sometimes spoil otherwise perfect chronological progressions in their diagrams of, say, entire national literatures – now serves to emphasize the telling neighbourhood of Starnone and Ferrante. The other main suspect, Raja, is nowhere close, attached, as she is (medium grey), to another extremity of the corpus. It is equally interesting, and equally telling, that Starnone's three texts written before the appearance of the Ferrante persona lie between the main corpus and the later Starnone/Ferrante nebula, thus further emphasizing the similarity of the Starnone and the Ferrante hands at a time of their shared "existence".

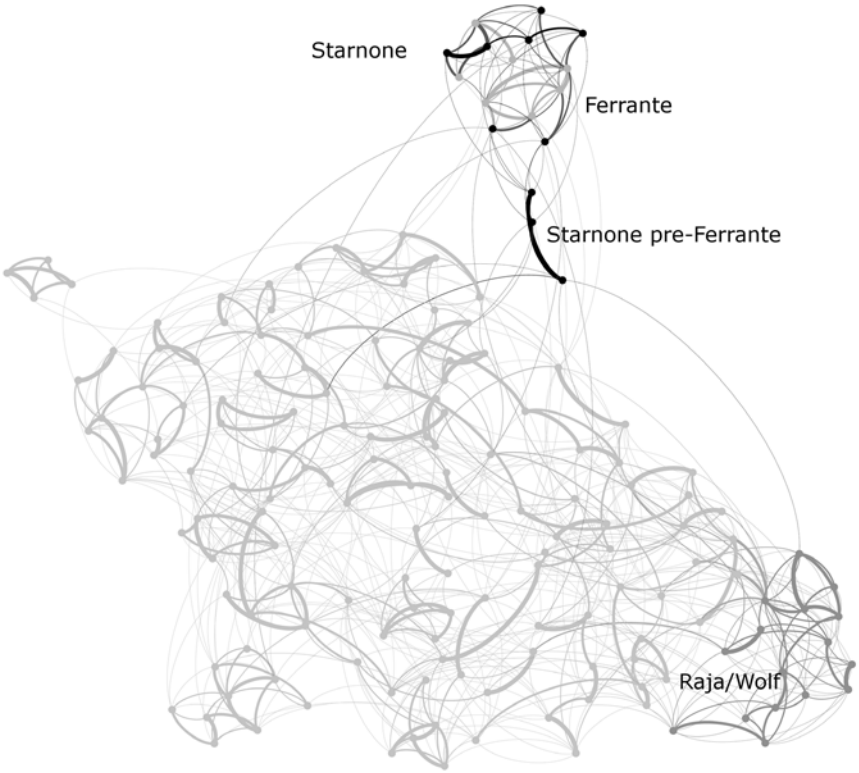

Figure 3. Starnone, Ferrante and Raja/Wolf in a network of 150 texts by Italian "suspects".

Obviously, the above-mentioned Gatti investigation has placed the focus very strongly on just two suspects, and his insistence on Raja rather than Starnone certainly deserves serious consideration. This is why, apart from the above general study, I compared texts by Ferrante, Starnone and Raja's translations of Wolf in a separate series of tests. Figure 4 shows a network visualization of this set, and results are quite clear again.

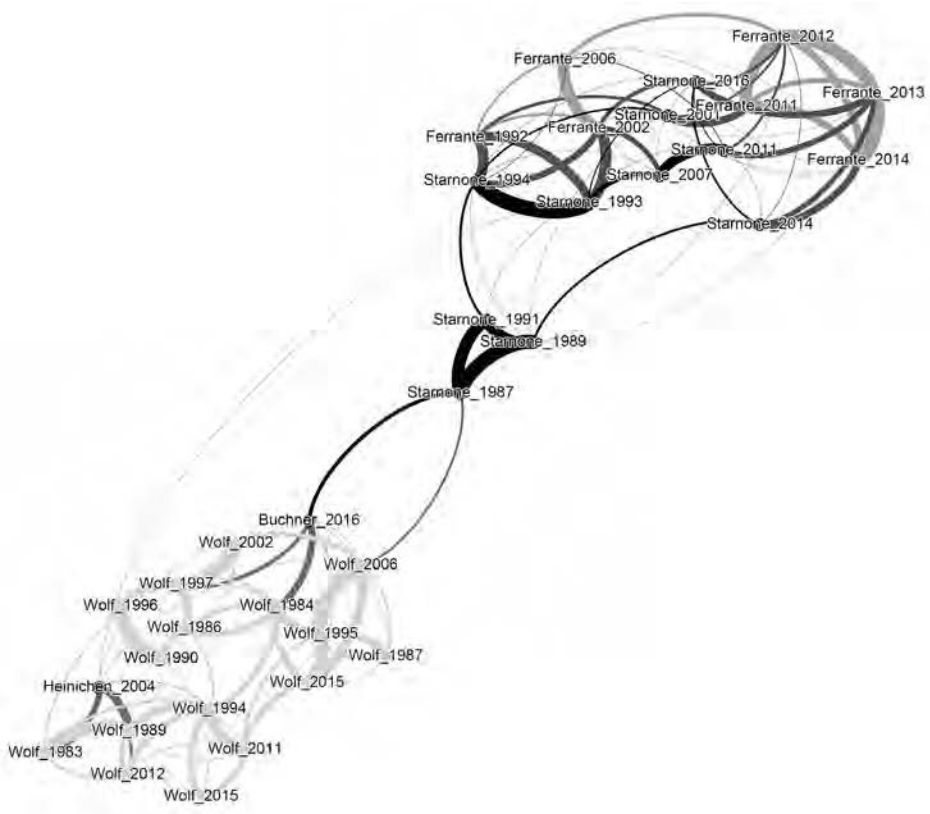

Figure 4. Network visualization of similarities between texts by Ferrante, Starnone and Italian translations of Wolf.

Here, too, Starnone seems to be married to Ferrante rather than to Raja; all of his texts (black) are interposed between the works by Ferrante (dark grey) and Italian renditions of Wolf by his translator wife (light grey); the same cluster contains two translations of Wolf by other people. Noteworthy is the strength of the links between texts by Starnone and Ferrante, and the virtual absence of any trace of linkage between Ferrante and Wolf/Raja; again, early Starnones form a separate subcluster, while his work in the post-Ferrante era is right next to "her" texts. This pattern is reiterated in a cluster analysis consensus tree in

Fig. 5, which is divided into two major branches: Wolf is on one, Starnone and Ferrante share the other.

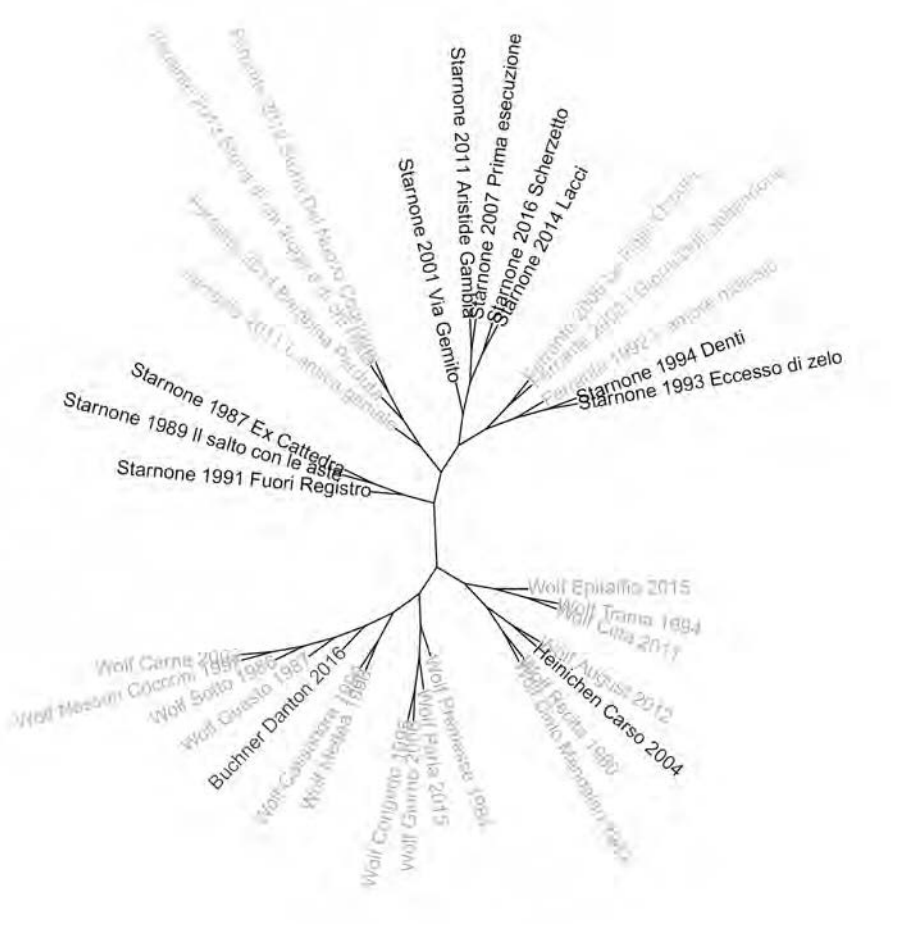

Figure 5. Cluster analysis consensus tree for texts by Ferrante, Starnone and Italian translations of Wolf.

The same conclusions are suggested when individual distance values between texts by Ferrante and those by Starnone and Raja are compared directly rather than through cluster analysis. Figure 6 shows individual Cosine Delta scores for each Ferrante book. They are invariably the lowest for novels by Starnone (various shades of grey). Not only do they come first: the scores for Wolf/Raja texts (black) come seventh at best.

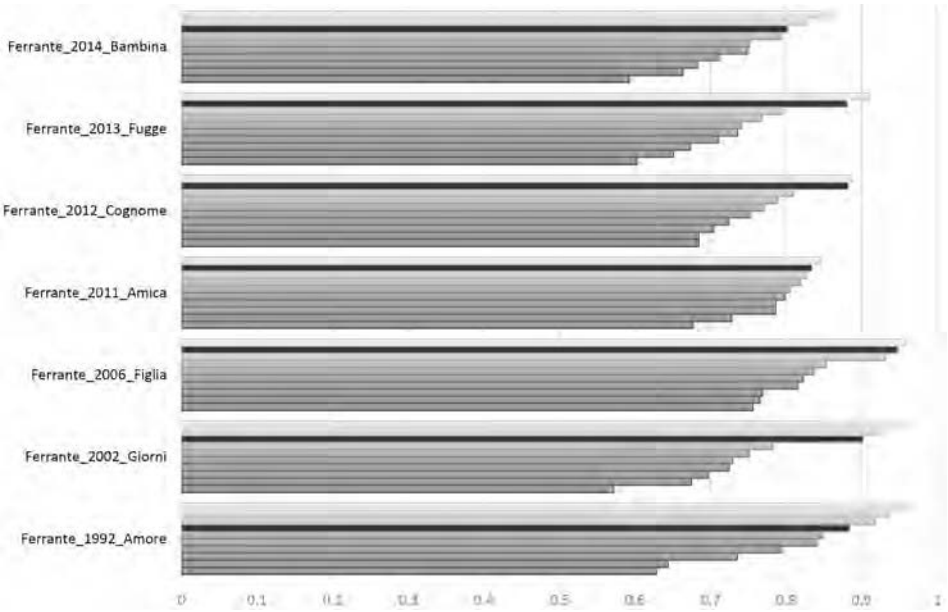

Figure 6. Delta scores between individual books in the collection.

Discussion

Stylometric evidence is very, very strong: the novels by Elena Ferrante have in fact been written by Domenico Starnone. Either that, or most-frequent-word based authorship attribution should never have been be trusted and never be trusted in the future. But this is humanities, and in the humanities, more than anywhere else, all kinds of things happen that were not dreamt of in our statistics.

First of all, this is not the first time that I investigated a collective creative writing effort by a couple. In a recent study, I looked for the respective strengths of the authorial signal of Jacek Dehnel, a writer, and his partner Piotr Tarczyński, a translator, in their two joint novels recently published in Poland. Contrarily to the Starnone/Raja ménage, the Polish authors never denied having written the two texts; quite to the contrary, they always presented that as their joint effort. And yet there, too, the stylometric signal in the texts was exclusively that of the writer Dehnel – as if Tarczyński had no hand in this, and I know he had (2016). There is not enough evidence to speak of a universal mechanism in such cases; but this is one thing that might bring Anita Raja back into the Ferrante story.

Secondly, another important caveat has to be made as to the very conditions of my experiment: there is no running away from the unwelcome fact that Raja's stylometric signal was modelled on that of her *translations* of a *single* writer, Christa Wolf, rather than her *own* writing. Stylometric studies into this area present a mixed picture: on the one hand, in translations of the same text, or even of the same original author, made by different people, the differences of the translators' own signals can be discernible (Rybicki and Heydel, 2013). On the other, when more original authors and more translators are involved, some translators seem to have their own stylometric fingerprint, while others successfully avoid identification (Burrows, 2002). Very often, the translated texts cluster by the original author (Rybicki, 2012, 2016). Raja may belong to either of those translators, but – especially since she has only translated Wolf – there is no way to find out which. From this point of view, the entire experiment would contain an inherent flaw, and this raises important doubts. Obviously, stylometric authorship attribution of the kind performed in this study is quite helpless if the real author is not present in the reference set of texts, and Raja might in fact be absent, and all I was comparing with Ferrante was some sort of an Italian stylometric signal for Wolf. Even this, however, cannot deny the fact that Ferrante and Starnone appear as a *separate* group *outside* the main network in Fig. 3.

This, in turn, continues to go against the evidence of many readers of Ferrante. They are quite adamant that, to write what she writes, she must be a woman. This must not be shrugged off as an "intuitive" fallacy; after all, it is shared by literary professors, her translators and her publishers, i.e. different categories of very reliable readers. What is more, their evidence does not necessarily quarrel against that of stylometry. It would be quite plausible that, in the light of all of my three caveats, Ferrante is in fact a joint effort of Raja and Starnone. Further speculation might be made that Raja is providing the content and Starnone is the one who clothes it in words, or that there is a less-identifiable – and thus an even more interesting – mode of literary collaboration. Raja's stylometric invisibility might also be somehow associated with her main profession; and that despite the fact that the common myth of translator's invisibility was famously deconstructed by Venuti (1995).

If not, and if it is, after all, solely, or mainly, Starnone, the entire question, "Who is Ferrante?" might in fact be turned around into, "Who is Starnone?" After all, this would mean that his own output, the one signed "Starnone," is, at least now, clearly less successful – in terms of sales and/or the number of foreign translations – than his "clandestine" creative activity as Ferrante. At a later time in the future, will he be remembered as the author of the famous *L'amica geniale* rather than of anything he wrote under his own name? Which, in fact, is *his* true literary identity? Or even: is there one? Let us not forget the evidence

of the three pre-Ferrante novels by Starnone: it might be telling us that Raja is just as present in Starnone as she is in Ferrante.

And that, perhaps, would be the nicest, or the most satisfying, answer to our question. In my private opinion, the sensational question of Ferrante's identity – although it may continue to make headlines in international press – is infinitely less important than that of the creative act of writing novels: what happens when one writes trying to write like someone else, or when one collaborates in that creative act with another person. Is it too much to ask that, one day, the real persons behind the Ferrante persona join forces with stymied scholars to advance our knowledge of literary creativity and creation?

References

Bastian, M., Heymann, S. and Jacomy, M. (2009). Gephi: an open source software for exploring and manipulating networks. In *Proceedings of the International AAAI Conference on Weblogs and Social Media*, San Jose, Ca.

Burrows, J. F. (1987). *Computation into Criticism.* Oxford: Clarendon Press.

Burrows, J. F. (2002). The Englishing of Juvenal: Computational Stylistics and Translated Texts, *Style,* 36(4), 677-698.

Eder, M. (2016). Rolling stylometry, *Digital Scholarship in the Humanities*, 31(3), 457-469.

Eder, M. (2017). Visualization in stylometry: cluster analysis using networks, *Digital Scholarship in the Humanities*, 32(1), 50-64.

Eder, M. and Rybicki, J. (2016). Go Set A Watchman while we Kill the Mockingbird in Cold Blood, with Cats and Other People. In *Digital Humanities 2016: Conference Abstracts.* Kraków: Jagiellonian University and Pedagogical University, 184-186.

Eder, M., Rybicki, J. and Kestemont, M. (2016). Stylometry with R: A Package for Computational Text Analysis, *The R Journal,* 8(1), 107-121.

Evert, S., Proisl, T., Jannidis, F., Reger, I., Pielström, S., Schöch, C. and Vitt, T. (2017). Understanding and explaining Delta measures for authorship attribution, *Digital Scholarship in the Humanities*, 32 (suppl. 2), 4-16.

Gatti, C. (2016). Elena Ferrante, le «tracce» dell'autrice identificata, *Il Sole 24 Ore – Domenica*, Milano, 2 October 2016, 1-2.

Gladwin, A. G., Lavin M. J. and Look, D. M. (2017). Stylometry and collaborative authorship: Eddy, Lovecraft, and "The Loved Dead".

Jacomy, M., Venturini, T., Heymann, S. and Bastian M. (2008). "ForceAtlas", a Continuous Graph Layout Algorithm for Handy Network Visualization Designed for the Gephi Software, *PLoS ONE* (2008), 9(6) (e98679. doi:10.1371/journal.pone.0098679).

Kenny, A. (1982). *The Computation of Style. An Introduction to Statistics for Students of Literature and Humanities.* Oxford: Pergamon Press.

Kestemont, M., Moens S. and Deploige, J. (2015). Collaborative authorship in the twelfth century: A stylometric study of Hildegard of Bingen and Guibert of Gembloux, *Digital Scholarship in the Humanities* 30(2), 199-224.

Van der Ploeg, J. (2017). Digitaal ontmaskerd, *De Volksrant*, Dec. 23rd, 15-17.

R Core Team (2014). *R: A language and environment for statistical computing*, http://www.R-project.org/.

Rybicki, J. (2012). The great mystery of the (almost) invisible translator. In Oakes, M. P. and Meng Ji (eds.), *Quantitative Methods in Corpus-Based Translation Studies: A practical guide to descriptive translation research.* Amsterdam: John Benjamins, 231-248.

Rybicki, J. (2016). Czy stylometria literacka to nowa humanistyka?, *Nowa humanistyka: zajmowanie pozycji, negocjowanie autonomii.* Poznań: Uniwersytet im. Adama Mickiewicza, 3-4 Nov.

Rybicki, J, Hoover, D. and Kestemont, M. (2014). Collaborative authorship: Conrad, Ford and rolling delta, *Literary and Linguistic Computing,* 29(3), 422-431.

Rybicki, J. and Heydel, M. (2013). The stylistics and stylometry of collaborative translation: Woolf's Night and Day in Polish, *Literary and Linguistic Computing,* 28(4), 708-717.

Smith, P. W. H. and Aldridge, W. (2011). Improving Authorship Attribution: Optimizing Burrows' Delta Method, *Journal of Quantitative Linguistics,* 18(1), 63-88.

Tuzzi, A. and Cortelazzo, M.A. (2018). What is Elena Ferrante? A comparative analysis of a secretive bestselling Italian writer, *Digital Scholarship in the Humanities* (online first https://doi.org/10.1093/llc/fqx066).

Venuti, L. (2008). *The Translator's Invisibility: A History of Translation* (2nd ed.). Abingdon: Routledge.

Elena Ferrante Unmasked

Jacques Savoy
University of Neuchâtel, Switzerland

Abstract

Text categorization domain proposes many applications and a classical one is to determine the true author of a document, literary excerpt, threatening email, legal testimony, etc. Recently a tetralogy called *My Brilliant Friend* has been published under the pen name Elena Ferrante, first in Italian and then translated into several languages. Various names have been suggested as possible author (e.g., Milone, Parrella, Prisco, etc.). Based on a corpus of 150 contemporary Italian novels written by 40 authors, different well-known computer-based authorship attribution methods have been employed to answer the question "Who is behind Elena Ferrante?" To achieve this objective, the Delta method, grounded on the 100 to 2,000 most frequent tokens or lemmas, reaches the conclusion that Domenico Starnone is the true author behind Elena Ferrante's pseudonym. As a second attribution strategy, Labbé's approach (*k*-nearest neighbor) was applied on the entire vocabulary and confirms this finding. A deeper analysis confirms this finding by revealing examples of close lexical similarities between Domenico Starnone and Elena Ferrante

Introduction

The Italian novelist Elena Ferrante became famous not only in Italy but also worldwide after the translation of her books, and particularly *L'amica geniale* (2011), the first novel of the well-known *My Brilliant Friend* tetralogy (e.g., bestsellers in the US and in several Western European countries). Living in a Ferrante fever, the general public wants to know more about this pen name. However, her true identity is still unknown even if some journalists and liter-

ary scholars have suggested some possible names (e.g., Erri De Luca, Francesco Piccolo, Michele Prisco, Fabrizia Ramondino, ...). However, no detailed scientific study has investigated this question, except a recent short preliminary report (Cortelazzo *et al.*, 2016; now Cortelazzo *et al.*, 2018) suggesting that Domenico Starnone could be the real writer. With computer-based authorship attribution models and a large corpus of 150 contemporary Italian novels, this study will analyze her writings and reveal her true identity or, at least, reduce this uncertainty to a reduced number of possible names. Such questions are not new in literacy history as, for example, in the French literature with the Gary-Ajar interrogation in 1970-75 or after the publication of the crime novel *The Cuckoo's Calling* (2013) under the pen name R. Galbraith by J.K. Rowling (Juola, 2015). In the 17th century one can find more numerous attribution debates in which the most well-known are Shakespeare (Craig and Kinney, 2009), the *Federalist Papers* (Jockers and Witten, 2010; Savoy, 2013), or the Corneille-Molière debate (Labbé, 2009; Marusenko and Rodionova, 2010).

The precise context of this study is the closed-class attribution question assuming the following general and weak assumptions. First, only standard and approved attribution methods (that have been examined and used by several studies) can be used. An important attribution cannot be ground on a single new approach never tested on other corpora and for which all aspects have not been well assessed and studied. Second, the real writer is one of the proposed authors in the underlying corpus. Third, for each novel, no collaboration took place during the writing process. Of course, one can admit that a collaboration might exist to develop the novel's outline and to elaborate some figures or dialogues. But the writing itself is clearly produced by a single person. Four, the name associated to a novel is the true writer. Moreover, all novels published under this name were written by that person. Finally, our attribution schemes will only take into account the textual elements. No other meta-data information (e.g., the author should be a female, must have lived in Naples, ...) will be considered to propose an attribution.

Without having the final true answer or the reliability of a DNA test, can we discover the true identity of the author based on a novel? Certainly, the text length is of prime importance and in our context the entire novel is available. However, how can we extract the discriminative stylistic features from a text to be able to detect the true author? How can we analyze or compare text representations to determine the author or to derive a short list of possible candidates?

To reveal the identity of Elena Ferrante, Section 2 presents the state of the art while Section 3 describes the collection of contemporary Italian novels supporting this research. As various authorship attribution approaches have been

suggested, without one clearly dominating the field, two different models have been employed as explained in Section 4. Section 5 exposes the main results achieved by the two selected attribution models. Section 6 depicts a more detailed analysis supporting our findings. Finally, a conclusion draws our main findings and some limits of this study.

State of the Art

As with other text categorization tasks, an effective authorship attribution model must represent each text according to a set of selected stylistic features reflecting the difference between the possible authors. Second, an intertextual distance function or a classifier must be chosen to determine the true writer.

To achieve this, a first family of methods suggests defining an invariant stylistic measure (Holmes, 1998) reflecting the particular style of a given author and varying from one person to another. As possible solutions, different lexical richness measures or word distribution indicators have been proposed such as Yule's K measure, statistics related to the type-token ratio (TTR), as well as the average word length, or the mean sentence or word length. None of these measures has proven very satisfactory due, in part, to word distributions ruled by a large number of rare events (LNRE) (Baayen, 2008).

As a second framework, different multivariate models can be applied to project each document surrogate into a reduced space under the assumption that texts written by the same author should appear close together. Some of the main approaches applicable here are principal component analysis (PCA) (Binongo and Smith, 1999; Craig and Kinney, 2009), hierarchical clustering (Labbé, 2007; Cortelazzo *et al.*, 2016), or discriminant analysis (Jockers and Witten, 2010). As stylistic features, these approaches tend to employ the top 50 to 200 most frequent word-types (MFW), as well as some POS information.

As a third useful paradigm, and based on various word selection schemes, different distance-based measures have been suggested. As well-known strategies, one can mention Burrows' Delta (2002) using the top m most frequent words (with m = 40 to 1,000), the Kullback-Leibler divergence (Zhao and Zobel, 2007) using a predefined set of 363 English words, or Labbé's method (2007) using the entire vocabulary and opting for a variant of the Tanimoto distance.

As a forth family of models, various machine learning approaches have been suggested (Abbasi and Chen, 2008; Stamatatos, 2009; Jockers and Witten, 2010) as, for example, decision trees, back-propagation neural networks, k-NN, and support vector machines (SVM), the latter being a popular approach in various CLEF campaigns (Stamatatos *et al.*, 2015). Zheng *et al.* (2006) found that SVM and neural networks tended to produce similar performance levels that are

significantly better than those achieved by decision trees. The *k*-NN approach tended to produce better effectiveness than both the Naïve Bayes or decision tree (Zhao and Zobel, 2007). Jockers and Witten (2010) showed that the Delta scheme could surpass performance levels achieved by the SVM method.

Finally, if words seem a natural way to generate a text surrogate, other studies have suggested using the letter occurrence frequencies (Kjell, 1994) or the distribution of short sequences of letters (character *n*-grams) (Juola, 2008). As demonstrated by Kešelj *et al.* (2003), such a representation can produce high performance levels. When adopting such a strategy, the final decision is more difficult to clearly explain to the user (e.g., what is the stylistic element or meaning of "ui"?). Finally, the fingerprint of an author can also be identified by the POS tags distribution of short sequences of such tags. Such text representations do not usually produce the best performance levels, but can be used as useful complementary information (Stamatatos *et al.*, 2015).

Corpus

To solve the Ferrante mystery, a team of researchers at the University of Padova under the supervision of Prof. Arjuna Tuzzi and Prof. Michele Cortelazzo has generated a corpus of contemporary Italian novels called PIC (Padova Italian Corpus). The list of authors appearing in this collection is given in the Appendix together with their gender and, for some novels, their length and publication year. This collection contains 150 books dedicated to adult readers and written by 40 different authors (27 men, 12 women, and Ferrante). Each novelist appears with at least two works, with a maximum of ten (Starnone). This corpus includes seven novels authored by Ferrante (including her well-known tetralogy). A careful editing process has been applied to remove all elements not belonging in the text itself (e.g., page number, running titles, etc.) as well as a thorough control of the spelling.

In selecting the authors, all suspected novelists behind the pseudonym Ferrante have been included. Thus, novels written by ten authors from Naples and the region of Campania are incorporated (e.g., De Luca, De Silva, Milone, Montesano, Parrella, Piccolo, Prisco, Ramondino, Rea, Starnone). This aspect could be important in Italian due to the presence of some spelling differences across regions (diatopic variation), and the use of dialectic expressions. Moreover, the corpus also contains bestsellers and awarded works. With four exceptions, the novels have been published from 1987 to 2016.

In total, the corpus contains around ten million word tokens (9,609,234) with an average of 64,062 tokens / novel (standard deviation: 38,228). The smallest text is composed of 7,694 tokens (Parrella, *Behave*, 2011) and the largest

196,914 tokens (Faletti, *Io uccido*, 2002). Only one novel contains less than 10,000 word-tokens. When considering only Ferrante's novels, one can see that their average size is 88,933 word-tokens (min: 36,222 (*La figlia oscura*), max: 138,622 (*Storia della bambina perduta*)), an average larger than the corpus mean (64,062 tokens). All of Ferrante's writings represent 6.48% of the corpus, while Faletti's books constitutes the largest part (6.6%) followed by Starnone (6.4%), and Mazzucco (6.15%). The smallest contribution is provided by Parrella (0.36%), followed by Vinci (0.58%), then Nori (0.64%). More information about this corpus can be found in Tuzzi and Cortelazzo (2018).

As preprocessing for all experiments, the text of each novel has been analyzed by the TreeTagger POS tagger (Schmid, 1994) to derive both the word-tokens and the lemmas (dictionary entries). Then all uppercase letters are transformed to their lowercase equivalents and all punctuation symbols or digits have been removed. This decision is grounded on the fact that the punctuation symbols can be present in different visual forms on the one hand, and on the other, those punctuation symbols can be imposed or modified by the editor.

In conclusion, from a computer-based attribution perspective, it is important to underline that this corpus possesses two essential characteristics. Each text contains more than 10,000 word-tokens on the one hand, and, on the other, a rigorous spelling control process has been applied.

Authorship Attribution Models

To solve the Ferrante mystery, we propose to rely not on a single attribution model but to consider several approved attribution methods grounded on different sets of features. First, the Delta approach (Burrows, 2002; Evert *et al.*, 2017) is selected, an approach based on the most frequent word-types (MFT) or lemmas (a set composed mainly of function words such as determiners, pronouns, prepositions, conjunctions, and some auxiliary verb forms). As feature set size, a value between 100 to 400 MFT is the norm, words defined without Ferrante's novels. To weight each term (word-token or lemma), we do not directly take account of the relative or absolute frequency, but rather their standardized frequencies (Z score). Such a value is obtained by subtracting the mean and dividing by the standard deviation. More precisely, for each term t_i in a corpus, its relative term frequency rtf_{ij} in a text T_j is computed as well as the mean ($mean_i$), and standard deviation (s_i) of that term over all novels belonging to the corpus (see Equation 1).

$$Z\ \text{score}(t_{ij}) = (rfr_{ij} - mean_i) / s_i \qquad (1)$$

Given a query text Q, an author profile A_k (concatenation of all his/her writings), and a set of terms t_i, for $i = 1, 2, \ldots, m$, the Delta value is computed according to Equation 2. Large differences may occur when, for a given term, both Z scores are large and have opposite signs. In this case, one author tends to use the underlying term more frequently than the mean while the other employs it rarely. When for all terms the Z score values are very similar, the distance between the two texts is small.

$$\Delta(Q, A_k) = {}^{1}/_{m} \cdot \sum_{i=1}^{m} \left| \text{Z score}(t_{iq}) - \text{Z score}(t_{ik}) \right| \quad (2)$$

As a second attribution model, the intertextual distance proposed by Labbé (2007) has been applied. This function returns a value between 0 and 1 depending on the degree of overlapping between the two texts. A value of 0 indicates that the two texts are identical, using the same vocabulary with the same frequencies for all terms. A distance of 1 specifies that the two novels have nothing in common (e.g., one in Italian, the other in Finnish). Between these two limits, the returned value depends on the number of words appearing in both texts and their occurrence frequencies.

More formally, the distance between the Texts A and B (denoted D(A,B)) is depicted by Equation 3 where n_A indicates the length of Text A (number of tokens), and tf_{iA} denotes the absolute term frequency of word-type i (for $i = 1, 2, \ldots, m$). The length of the vocabulary is indicated by m. Usually Text B does not have the same length (here it is assumed that its length is larger than Text A). To reduce the longest text to the size of the smallest, each of term frequencies (tf_{iB}) is multiplied by the ratio of the two text lengths as indicated in the second part of Equation 3.

$$D(A, B) = \sum_{i=1}^{m} \left| tf_{iA} - \widehat{tf}_{iB} \right| / (2 \cdot n_A) \quad \text{with } \widehat{tf}_{iB} = tf_{iB} \cdot {}^{n_A}/_{n_B} \quad (3)$$

Identification of Elena Ferrante's True Identity

The two selected attribution models have been applied to the PIC corpus in which the seven novels written by Elena Ferrante form the test set, and the rest the training set.

The Delta Method

Using the Delta model, all novels written by an author are concatenated to generate the corresponding author profile. For each of the seven Ferrante's novels, this model has been applied with, as feature set, the 100, 200, 300, 400, 500,

1000, or 2000 most frequent word-types (MFT) or lemmas (MFL). Usually, values between 200 and 500 are effective in identifying the true author of a document or text excerpt (Savoy 2015). Those sizes also correspond to values indicated in the basic paper on this approach (Burrows 2002). For all these 7 (novels) x 7 (feature sets) x 2 (tokens or lemmas) = 98 experiments, the same name appears in the first rank: Domenico Starnone.

To have a more precise view of this attribution method, Table 1 indicates the top five names sorted by the Delta model using the 200 MFL with two Ferrante's novels, namely *L'amore molesto* (her first novel, published in 1992), and *L'amica geniale* (the first book of her tetralogy, 2011).

Table 1. Ranked list of possible authors according to the Delta method (200 lemmas).

	L'amore molesto		*L'amica geniale*	
Rank	Distance	Author	Distance	Author
1	0.634	Starnone	0.481	Starnone
2	0.824	Brizzi	0.689	Veronesi
3	0.830	Giordano	0.691	Balzano
4	0.842	Lagioia	0.731	Nesi
5	0.851	Milone	0.733	Brizzi
…	…	…	…	…

It is interesting to note that the distance value difference between the first and the second author is larger compared to the difference between the second and third. For example, the difference between the first two ranks with *L'amore molesto* is 0.824 − 0.634 = 0.19 (or 30%). The gap between the second and the third is 0.830 − 0.824 = 0.006 (or 0.7%). This comparison indicates that the first answer is clearly more probable than the other novelists in the ranked list. A similar finding can be found with the second novel depicted in Table 1.

When using word-tokens as features, the following names appear in the second rank in our 49 experiments (7 (novels) x (feature sets)), namely Veronesi (23 times), Milone (9), Brizzi (6), Tamaro (5), Sereni (2), Carofiglio (2), Balzano (1), and Mazzucco (1). Applied the Delta method with lemmas, the following novelist appears in the second rank: Veronesi (27 times), Giordano (8), Milone (4), Brizzi (4), Sereni (4), Balzano (1), and Maraini (1).

The Labbé's Intertextual Distance

With Labbé's distance-based model (Labbé, 2007), instead of using the word-tokens as done previously, the lemmas have been used. As the Italian language owns a richer morphology compared to English, the lemma can reduce some variations present in the tokens and judged ineffective in determining the style (e.g., from the tokens *amico*, *amica*, *amici*, the same lemma (*amico*,

friend) is derived). To determine these dictionary entry forms, the TreeTagger POS tagger has been applied. When generating each text representation, the lemmas having an occurrence frequency of one or two have been ignored. The intertextual distance was then computed for all novel pairs and ranked from the smallest distance to the largest one. Table 2 presents a fragment of the returned output and the full title of the corresponding novels can be found in Table A.2 in the Appendix.

As reported in Table 2, the smallest distances are always associated with works written by the same author. As soon as the distance value increases, our certainty that both texts were written by the same person decreases. In Rank #33, the first "incorrect" pairing is discovered, and other similar examples occur in Position #38, #41, and #42. Looking below, one can find eleven additional "incorrect" pairings up to Rank #84 where the first (real?) erroneous link occurs. This result indicates that the real author behind Elena Ferrante's writings is certainly Domenico Starnone. Having more than ten "incorrect" assignments between these two names before another pairing appears is clearly surprising. Moreover, the distance for the first one (0.193) is quite small for the possibility of two distinct authors.

Table 2. Ranked list of novel pairs using Labbé's distance with lemmas.

Rank	Distance	DocID	Author	DocID	Author
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
...	...	...	...	...	...
33	0.193	52	Ferrante	132	Starnone
38	0.195	51	Ferrante	131	Starnone
41	0.196	51	Ferrante	132	Starnone
42	0.196	47	Ferrante	127	Starnone
...	...	...	...	...	...
84	0.216	25	Carofiglio	147	Veronesi

The next step is to estimate the probability that such a small distance value (0.193) can be found only between two texts written by the same person. To define this probability (Savoy, 2016), one can model the distance values (some examples are given in Table 2) as derived from a mixture of two Gamma distributions, one between two novels written by the same person (distribution D1

in Figure 1), the second with pairs linking papers produced by two distinct persons (distribution D2). The choice of the Gamma distribution can be explained by considering the fact that the distance values are never negative, can take all positive values (e.g., when using the Canberra distance function instead of Tanimoto or Labbé), and are skewed on the left.

According to the formalism described in (Savoy, 2016), one can estimate the probability that Starnone is the author of *Storia della bambina perduta* (DocID = 52 in Table 2) with a probability of 0.98. In fact, with texts longer than 10,000 words and belonging to the same genre, observing a distance value lower than 0.2 is a very strong indication that the same person wrote the two novels (Labbé, 2007).

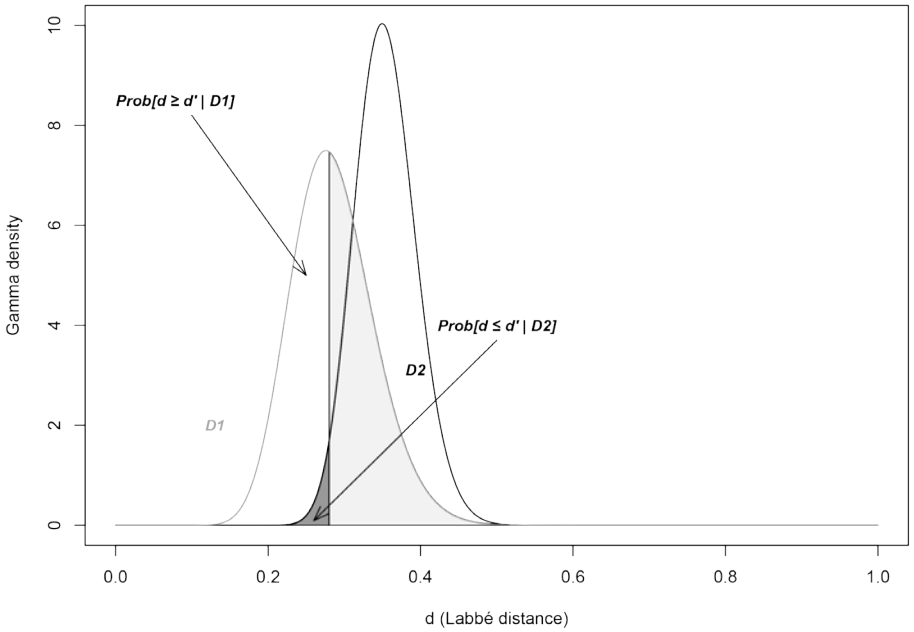

Fig. 1. Mixture distribution model with distance values between papers written by the same author (D1) on the left, and on the right (larger distance values) between texts written by two distinct authors (D2).

As a variant, instead of considering all novels separately, the corresponding writer profiles can be generated using the lemmas occurring in all their novels. As previously, an intertextual distance is computed between all 40 author profiles. The smallest distance (0.177) is observed between Ferrante and Starnone's

profile. The second smallest (0.22) appears between Piccolo and Veronesi, the third (0.226) between Nesi and Veronesi, and the forth (0.227) between De Silva and Veronesi. The distance gaps between the second, third, and forth are all rather small (0.007, 0.006, 0.01) compared to the interval between the first and the second (0.22 – 0.177 = 0.043) confirming the very close lexical proximity between Ferrante and Starnone. The conclusion is similar when adopting the word-types instead of the lemmas.

Deeper Analysis

A deeper analysis reveals several reasons explaining the strong lexical similarity between Starnone and Ferrante discovered by the two attribution models. First, the focus will be set on frequent words, knowing that Starnone's novels represents 6.4% (615,238 / 9,609,234) of the corpus and Ferrante 6.48% (622,532 / 9,609,234) (see Table 3). Compared to all other novelists, the word-type *padre* (father) occurs in total 9,815 in the corpus, but proportionally more frequently in Ferrante's novels (833 occurrences, 8.5%) or in Starnone's writings (1,170, 11.9%).

Table 3. Distribution of the word "padre" in Ferrante, Starnone, and other novelists present in our corpus.

	Ferrante	Starnone	Others	Total
"padre"	833	1,170	7,812	9,815
Other words	621,699	614,068	8,363,652	9,599,419
Total	622,532	615,238	8,371,464	9,609,234

As Ferrante represents 6.48% of the total, we would expect having 636 occurrences of the word *padre* in her writings (0.065 x 9,815). However, we observe 833 occurrences, clearly a larger amount. Similarly with Starnone, we would expect 628 times the word *padre* (0.064 x 9,815) instead of 1,179 observed ones. Finally, for the remaining 38 authors, we see 7,812 occurrences while the expected number is 8,551 (0.87 x 9,815). Those differences form the basic information for the chi-square test applied to verify whether the word distribution differs significantly across the authors (with a significance level of 0.1%) (Oakes and Farrow, 2007).

Similar distributions can be observed with the word *madre* (mother) having a frequency of 8,246 in the corpus, 1,104 in Ferrante's (13.4%), and 762 in Starnone's works (9.2%). Additional examples can be found and Table 4 reports other word-types such as *perciò* (therefore) occurring 1,263 times in the entire

corpus, with 222 occurrences in Ferrante's novels, and 254 in Starnone's books. In the last column of Table 4, the chi-square test has been applied (with a significance level of 0.1%).

On the other hand, some word-types are employed only by these two writers, such as *contraddittoriamente* (contradictory), *giravite* (screwdriver), *studenti* (students), *soffertamente* (by suffering). An interesting example is the word-type *malodore* (stink) appearing as this spelling in both Ferrante and Starnone's novels but the same meaning could appear as *maleodore*. This last spelling appears in other novels but never under the pen of Ferrante or Starnone.

Table 4. List of words occurring more frequently in Ferrante and Starnone's novels.

Word	Corpus	Ferrante	Starnone	Significant?
padre (father)	9,815	833	1,170	yes
madre (mother)	8,246	1,104	762	yes
perciò (therefore)	1,263	222	254	yes
temere (fear)	1,345	274	207	yes
persino (even)	1,351	266	205	yes
tono (tone)	2,135	421	286	yes
gridare (shout)	2,201	399	303	yes
mostrare (to show)	2,271	384	310	yes
contento (happy)	1,665	280	227	yes
brutto (ugly)	1,893	327	243	yes
frase (phrase)	2,182	334	312	yes

As a third strata of word frequency, one can consider word-types showing a low occurrence frequency, and more precisely, those that occur more often in Starnone's and Ferrante's books compared to the other Italian authors. For example, the term *minutamente* (minutely) occurs 28 times in Ferrante's novels, 14 times in Starnone's writings, and three times in the rest. With *tassare* (to tax), one can observe something similar; 22 with Ferrante, 10 with Starnone, three times for the others. The word-type *reattività* (reactivity) occurs 22 times in the whole corpus, and Ferrante employs it six times and Starnone 13 times. Our last example is related to dialect usage with the word *strunz* (shit). This term does not belong to the classical Italian language (in which it is spelled as *stronzo*) but corresponds to a Neapolitan dialect form. The occurrence distribution for this word is the following: 18 times in Ferrante's novels, 63 times in Starnone's writings, and four times for all the others (two times in De Silva's, and two times in Raimo's novels).

As another way to analyze the lexical proximity between the novels written by Ferrante and Starnone, a variant of the Zeta approach suggested by (Burrows

2007) can be applied (Craig and Kinney 2009). In a first stage, the terms appearing recurrently in text passages written by Author A and rarely in excerpts written by another writer (denoted B) must be defined. The objective consists of tracking the presences and absences of the words, not their occurrence frequency. To achieve this, all novels are decomposed into non-overlapping chunks of size k (e.g., k = 4,000 in the current study). In the following, we assume that n_{Ak} chunks have been extracted from novels written by A and n_{Bk} chunks from books written by the other novelist.

Second, a discriminate weight w_i is assigned to each term t_i according to Equation 4. A higher weight is given to words appearing more often in passages written by A than in those written by B. To achieve this, df_{Ai} denotes the number of chunks written by A having at least one occurrence of the term t_i (and similarly for df_{Bi}). According to Equation 4 representing the summation of two proportions, each term t_i will have a value between 0 and 2.

$$w_i = {df_{Ai}}/{n_{Ak}} + {(n_{Bk} - df_{Bi})}/{n_{Bk}} \quad (4)$$

When a term occurs in all chunks written by A ($df_{Ai} = n_{Ak}$), and never in B ($df_{Bi} = 0$), w_i reaches the maximum value of 1 + 1 = 2. On the other hand, when the word occurs in all passages written by B ($df_{Bi} = n_{Bk}$), and never with A ($df_{Ai} = 0$), the discriminative value w_i is 0 + 0 = 0. Finally, if the word is very frequent and appears in all chunks ($df_{Ai} = n_{Ak}$ and $df_{Bi} = n_{Bk}$), the resulting weight is 1 + 0 = 1.

Third, sorting the words according to decreasing discriminative weights, the terms associated with Author A appear on the top, with values larger than 1.0. On the bottom part, one can identify words ignored (or used rarely) by A and occurring frequently with the novelist B. In this study, we selected the top 750 words having a discriminative weight larger than 1 to form the vocabulary specific to A and the lower 750 words (with a weight smaller than 1) to determine the words associated with B.

Finally, based on these two word lists, one can visualize the lexical proximity between a given novel (composed of a set of passages) and both Authors A and B. To achieve this, the studied text is divided into non-overlapping chunks (of size k). For each passage, the percentage of words appearing in both lists indicates the two coordinates. Similarly, texts written by A and B can be decomposed and each passage can be added into the graph.

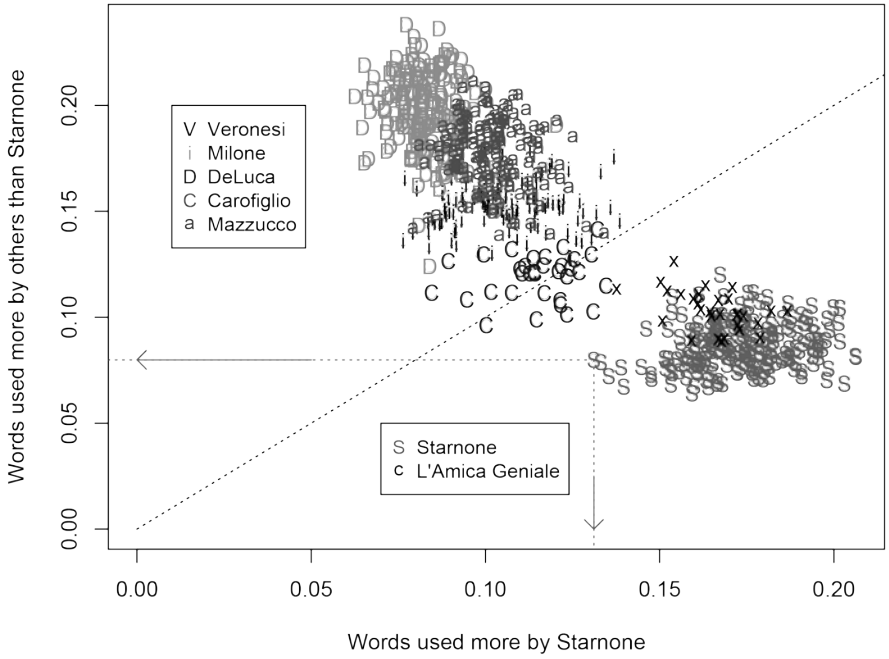

Fig. 2. The novel *L'amica geniale* compared to books written by Starnone, Carofiglio, De Luca, Mazzucco, Milone, and Veronesi.

In Figure 2, the 750 words more specific to Starnone are used to define the X-coordinate while the terms appearing more frequently in novels written by Carofiglio, De Luca, Mazzucco, Milone, and Veronesi define the Y-coordinate. The tested novel is *L'amica geniale* (Ferrante) subdivided into 29 passages (of 4,000 words). As depicted in Figure 2, all these points appear in or very closed to Starnone's cloud. For a passage written by Starnone, the precise coordinates are indicated by an array (x-value: 23.7%; y-value: 13.5%).

This deeper analysis reveals the close lexical proximity that can be found between Starnone and Ferrante. With both topical terms (e.g., *padre* (father), *madre* (mother), *temere* (fear), *giravite* (screwdriver)), and functional words (e.g., *persino* (even), *perciò* (therefore)), the occurrence distributions are similar for both Starnone and Ferrante and different from the other novelists. Applying this variant of the Zeta test (Burrows, 2007) confirms the conclusion of the two well-known authorship attributions.

7. Conclusion

The two standard and approved attribution models reach the same conclusion: Domenico Starnone is the hidden hand behind Elena Ferrante. Varying the parameter values of the two text categorization approaches does not change this conclusion. Whether considering tokens or lemmas as features, the same result always appears. Modifying the feature set size does not modify this finding. Applying a classifier grounded on author profile or one that is instance-based produces the same overall attribution. Thus, considering their lexical proximity, all methods indicate the same name behind Elena Ferrante's novels.

The underlying corpus contains all novelists that have been mentioned as possible secret hands behind Ferrante. This set contains ten authors originating from the region (Campania) that appear in the background of the *My Brilliant Friend* tetralogy. In addition, when generating this corpus, thirteen female writers have been selected. Therefore, one can conclude that a real effort has been deployed to include many authors sharing some important extra-textual relationships with Ferrante (e.g., a woman coming from Naples or environs).

We must however acknowledge that a collaboration between two (or more) persons might exist, for example, to draw some psychological traits of figures appearing in the novels, to elaborate part of the scenario, or to imagine some replicas of a dialogue. Nevertheless, according to our study, the writing process is the fruit of a single person.

This conclusion is reached under the *closed-set* hypothesis, assuming that the real author is one of the 39 proposed novelists. Is it possible that another unknown writer is the true author of all Ferrante's books? Under this *open-set* hypothesis, our conclusion might be wrong because we cannot exclude, with 100% certainty, that no other person is behind Ferrante. However, when applying the Labbé's intertextual distance, the distance found in Rank #33 (see Table 2) between one Ferrante's novel (*Storia della bambina perduta*) and one Starnone's book (*Lacci*) is very small (0.193). In fact, such a small value (smaller than 0.2) never appears between documents written by two distinct authors, when considering text sizes longer than 10,000 word-tokens and written in the same genre and time period (Labbé, 2007). Moreover, the probability associated with this assignment is very high (98%), indicating a strong evidence that Starnone wrote both books. In addition, the distance distribution depicted by the Delta method (see Table 1) is also a strong indication that Starnone is the true author. The distance difference between all other possible writers are rather small compared to the difference between Starnone and the possible second author.

As meta-information, many persons are convinced that Elena Ferrante must be a woman. To this prior argument, only an exceptional writer can produce a novel that is perceived as written by a person from the other gender (e.g.,

Joyce is an example, and now Starnone another). In Joyce's case, a study has demonstrated that inside the novel *Ulysses*, text passages corresponding to a dialogue between men or between women can be clearly attributed either to a male or female author. But we know that both were written by Joyce. Finally, we can mention that Domenico Starnone himself does not corroborate our conclusion (Fontana 2017), and the mystery about the name Ferrante remains; it could be related to Elsa Morante (1912-1985), a supposed ghostwriter for A. Moravia (1907-1990), her husband.

References

Abbasi, A. and Chen, H. (2008). Writeprints: A Stylometric Approach to Identity-level Identification and Similarity Detection in Cyberspace. *ACM – Transactions on Information Systems*, 26(2), 7.

Baayen, H.R. (2008). *Analysis Linguistic Data: A Practical Introduction to Statistics using R*. Cambridge: Cambridge University Press.

Binongo, J.N.G. and Smith, M.W. (1999). The Application of Principal Component Analysis to Stylometry. *Literary and Linguistic Computing,* 14(4), 445-465.

Burrows, J.F. (2007). All the Way Through: Testing for Authorship in Different Frequency Stata. *Literary and Linguistic Computing*, 22(1), 27-47.

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

Cortelazzo, M.A., Nadalutti, P., Ondelli, S. and Tuzzi, A. (2016). Authorship Attribution and Text Clustering for Contemporary Italian Novels. In Abstracts *Qualico2016*, EAQL, Trier, 7-8.

Cortelazzo, M. A., Nadalutti, P., Ondelli, S. and Tuzzi, A. (2018). Authorship Attribution and Text Clustering in Contemporary Italian Novels: Does Elena Ferrante's and Domenico Starnone's regional origin play a role? In: Wang, L., Köhler, R., Tuzzi, A. (eds.), *Scructures, properties, and interrelations. Selected papers from Qualico 2016*. Lüdenscheidt: RAM Verlag, 1-14.

Craig, H. and Kinney, A.F. (2009). *Shakespeare, Computers, and the Mystery of Authorship*. Cambridge: Cambridge University Press.

Evert, S., Proisl, T., Jannidis, F., Reger, I., Pielström, S., Schöch, C. and Vitt, T. (2017). Understanding and Explaining Delta Measures for Authorship Attribution, *Digital Scholarship in the Humanities*, 32(2), ii4-ii16.

Fontana, E. (2017). Lo scrittore Domenico Starnone: "Io non sono Elena Ferrante", Il Giornale, 9 September 2017.

Holmes, D.I. (1998). The Evolution of Stylometry in Humanities Scholarship. *Literary and Linguistic Computing*, 13(3), 111-117.

Jockers, M.L. and Witten, D.M. (2010). A Comparative Study of Machine Learning Methods for Authorship Attribution, *Literary and Linguistic Computing*. 25(2), 215-223.

Juola, P. (2008). Authorship attribution, *Foundations and Trends® in Information Retrieval*, 1(3), 233-334.

Juola, P. (2015). The Rowling Case: A Proposed Standard Protocol for Authorship Questions, *Digital Scholarship in the Humanities*, 30(1), i100-i113.

Kešelj, V., Peng, F., Cercone, N. and Thomas, C. (2003). N-Gram-Based Author Profiles for Authorship Attribution. In *Proceedings of the Conference Pacific Association for Computational Linguistics*, 255-264, ACL.

Kjell, B. (1994). Authorship Determination Using Letter Pair Frequencies Features with Neural Networks Classifiers, *Literary and Linguistic Computing*, 9(2), 119-124.

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

Labbé, D. (2007). Experiments on Authorship Attribution by Intertextual Distance in English, *Journal of Quantitative Linguistics*, 14(1), 33-80.

Marusenko, M. and Rodionova, E. (2010). Mathematical Methods for Attributing Literary Works when Solving the "Corneille-Molière" Problem, *Journal of Quantitative Linguistics*, 17(1), 30-54.

Oakes, M.P. and Farrow, M. (2007). Use of the Chi-Squared Test to Examine Vocabulary Differences in English Language Corpora Representing Seven Different Countries, *Literary and Linguistic Computing*, 22(1), 85-99.

Savoy, J. (2013), *The Federalist Papers* Revisited: A Collaborative Attribution Scheme. In Proceedings ASIST, Montreal, November (2013).

Savoy, J. (2015). Comparative Evaluation of Term Selection Functions for Authorship Attribution, *Digital Scholarship in the Humanities*, 30(2), 246-261.

Savoy, J. (2016). Estimating the Probability of an Authorship Attribution. *Journal of the American Society for Information Science and Technology,* 67(6), 1462-1472.

Schmid, H. (1994). Probabilistic Part-of-Speech Tagging Using Decision Trees. In Proceedings *International Conference on New Methods in Language Processing*, Manchester.

Stamatatos, E. (2009). A Survey of Modern Authorship Attribution Methods, *Journal of the American Society for Information Science and Technology*, 60(3), 433-214.

Stamatatos, E., Potthast, M., Rangel, F., Rosso, P. and Stein, B. (2015). Overview of the PAN/CLEF 2015 Evaluation Lab. In Josiane Mothe et al. (eds.), *Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings*

6th International Conference of the CLEF Initiative (CLEF 15), 518-538, Berlin: Springer.

Tuzzi, A. and Cortelazzo M.A. (2018). What is Elena Ferrante? A comparative analysis of a secretive bestselling Italian writer, *Digital Scholarship in the Humanities* (online first 19 January 2018 fqx066, https://doi.org/10.1093/llc/fqx066).

Zhao, Y. and Zobel, J. (2007). Searching with Style: Authorship Attribution in Classic Literature. In Proceedings of the *Thirtieth Australasian Computer Science Conference*, 59-68, Ballarat.

Zheng, R., Li, J., Chen, H. and Huang, Z.A. (2006). A Framework for Authorship Identification of Online Messages: Writing-Style Features and Classification Techniques, *Journal of the American Society for Information Science and Technology*, 57(3), 378-393.

Appendix

Table A.1 reports the 40 author names included in the Padova Italian Corpus, with the corresponding gender (12 women, 27 men, one unknown (Ferrante)), and number of novels in the collection. The ten writers from Naples and the region of Campania are indicated in italics.

Table A.1. Author name, gender (M/F), and the number of novels in the PIC.

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

Table A.2. Some examples of novels included in the PIC corpus.

DocID	Author	Year	Size (tokens)	Title
25	Carofiglio	2010	69,645	Le perfezioni provvisorie
…	…	…		…
42	Faletti	2004	139,795	Niente di vero tranne gli occhi
44	Faletti	2009	127,363	Io sono dio
45	Faletti	2011	32,261	Tre atti e due tempi
46	Ferrante	1992	41,914	L'amore molesto
47	Ferrante	2002	53,546	I giorni dell abbandono
48	Ferrante	2006	36,222	La figlia oscura
49	Ferrante	2011	96,135	L'amica geniale
50	Ferrante	2012	138,622	Storia del nuovo cognome
51	Ferrante	2013	119,148	Storia di chi fugge e di chi resta
52	Ferrante	2014	136,945	Storia della bambina perduta
…	…	…		…
124	Starnone	1987	39,538	Ex cattedra
125	Starnone	1989	68,651	Il salto con le aste
126	Starnone	1991	38,031	Fuori registro
127	Starnone	1993	43,402	Eccesso di zelo
128	Starnone	1994	46,953	Denti
129	Starnone	2000	140,226	Via Gemito
130	Starnone	2007	40,787	Prima esecuzione
131	Starnone	2011	118,810	Autobiografia erotica di Aristide Gambia
132	Starnone	2014	36,554	Lacci
133	Starnone	2016	42,286	Scherzetto
…	…	…		…
145	Veronesi	2006	123,128,	Caos calmo
147	Veronesi	2007	105,722	Brucia Troia
…	…	…		…

Afterword

It is hardly by chance, we think, that the two sessions of the Workshop "Drawing Elena Ferrante's Profile" (Padua, 7 September 2017) were chaired by two literary scholars, the joint authors of this afterword, who look with interest at the methods of quantitative analysis of style and language in literary texts without ever actually having applied them (yet). Looking from a distance, and a distance that is growing increasingly shorter also for qualitative analysts, means acting as observers who experience many contrasting moods: utter wonder at the extensive coverage of corpora that only quantitative analysis provides, the degree of finesse that these multifarious methods have perfected and the nearly unanimous outcome that has sprung from this interaction between different methods applied to the same corpus, but also a slight sense of proprietary concern as to the possible implications of this triumph of quantitative methods.

"The machines have won" would be a crude way of putting it: the machines have not won yet. Even the most traditional scholar will sense here the presence of scholars who use quantitative methods as the best tools to tackle questions and techniques that would otherwise be left untried, not as proxy for humanistic commitment. Another thing that these studies make abundantly clear is that quantitative analysis seems at its best when it converses with traditional qualitative analysts and when a condition of mutual respect is ensured. And perhaps the very terms "quantitative" and "qualitative" need to be qualified: while the former tends to take on a fairly negative connotation at least among qualitative scholars, the latter could engender the wrong assumption that such an interpretation is based on an allegedly personal inclination towards literary interpretation fuelled by innate genius. There can be no quality without quantity: traditional assessments of literature are indeed based on endless readings.

Our roles as interested observers has been thus been compounded also with a composite sense of curiosity and some hint of concern, as it often happens when a new method and a new efficient way to tackle traditional problems gradually emerge within a scholarly discipline. Just like all academic fashions, traditional

literary analysis is not immune to change: only the rate of this change over time tends to be slower, and its new methods often borrow something from the previous ones, even (or especially) when they explicitly overturn them. Compared to this slow movement within academy, the practitioners of stylometry would instead risk being mistaken for a horde of aliens who not only do not need to actually read the works they are investigating (distant reading will do the trick for them), but can also be theoretically unfamiliar with the language itself of those works. Another unprecedented dimension of quantitative analysis that may baffle qualitative scholars is the maddening, incessant pace with which new methods are devised and improved radically in only a few months' time. This may be yet another manifestation of the vast penetration of computer science and its quick pace in all aspects of contemporary life, not only in academy. And, indeed, these studies also fuel other problems that need to be addressed: if a statistical method enables us to solve the authorship question regarding a work written under a pseudonym with very good approximation, can it be also legitimately used to probe into other questions, such as gender, age and political affiliations, that address privacy? The question is made quite topical by the general consensus reached through different ways by all of these scholars on the actual identity of Elena Ferrante as a man writing under a female pseudonym, perhaps in a still unclear degree of collaboration with a woman.

We still believe that there is more room for collaboration between quantitative and qualitative methods. Even more radically, we believe that traditional qualitative analysis can coexist with the quantitative one: the proximity between the works by Ferrante and only a select part of the corpus is a clear instance of a quantitative outcome that needs to be investigated by a literary scholar, addressing this stylistic gap as a starting point for further analysis and investigating whether it might have something to do with the outstanding international success of these works. And we also hope that quantitative scholars may fruitfully talk with their qualitative colleagues also *before* devising their studies, for many of the effects reported by quantitative methods often partially clarify insights that qualitative scholars already entertained thanks to their expertise. We also believe that distant reading may provide a fruitful detachment from the texts and a more general perspective within a genre or a literary field that escapes the single researcher's analysis, but which must then be followed by and mixed with a new close reading of the results, in a perpetual, virtuous circle.

It is comforting to note that somehow the literary author keeps moving away from analysis, be it quantitative or qualitative, and this explains the frequent references not only to standard methods of lexicometry (Ratinaud), data-compression (Lalli, Tria and Loreto) and thesaurus-based semantic similarity

(Juola), but also to images of movement (Eder's moving window), the metaphor of the mask (Savoy) and even of profiling (Mikros and Rybicki). Studying the works, not only those by the chimeric Elena Ferrante, literally requires "many hands" (Tuzzi and Cortelazzo). These studies prove that in the future these hands may also belong to quantitative and qualitative scholars who indulge in mutual dialogue.

Rocco Coronato and Luca Zuliani
University of Padova

Elena Ferrante is an internationally acclaimed Italian novelist whose real identity has been kept secret by E/O publishing house for more than 25 years. Owing to her popularity, major Italian and foreign newspapers have long tried to discover her real identity. However, only a few attempts have been made to foster a scientific debate on her work.

In 2016, Arjuna Tuzzi and Michele Cortelazzo led an Italian research team that conducted a preliminary study and collected a well-founded, large corpus of Italian novels comprising 150 works published in the last 30 years by 40 different authors. Moreover, they shared their data with a select group of international experts on authorship attribution, profiling, and analysis of textual data: Maciej Eder and Jan Rybicki (Poland), Patrick Juola (United States), Vittorio Loreto and his research team, Margherita Lalli and Francesca Tria (Italy), George Mikros (Greece), Pierre Ratinaud (France), and Jacques Savoy (Switzerland).

The chapters of this volume report the results of this endeavour that were first presented during the international workshop *Drawing Elena Ferrante's Profile* in Padua on 7 September 2017 as part of the 3rd IQLA-GIAT Summer School in *Quantitative Analysis of Textual Data*. The fascinating research findings suggest that Elena Ferrante's work definitely deserves "many hands" as well as an extensive effort to understand her distinct writing style and the reasons for her worldwide success.

978-88-6938-130-0